Exploratory analysis of indicators for open knowledge institutions: a case study of Australian universities

Author manuscript (not peer-reviewed)

Chun-Kai (Karl) Huang[1], Katie Wilson[1], Cameron Neylon[1,2], Alkim Ozaygen[1], Lucy Montgomery[1,2], Richard Hosking[1,2]
[1]*Centre for Culture and Technology, Curtin University, Bentley, WA 6102*
[2]*Curtin Institute for Computation, Curtin University, Bentley, WA 6102*

Abstract

While the movement for open access (OA) has gained momentum in recent years, there remain concerns about the broader commitment to openness in knowledge production and dissemination. Increasingly, universities are under pressure to transform themselves to engage with the wider community and to be more inclusive. *Open knowledge institutions* (OKIs) provide a framework that encourages universities to act with the principles of openness at their centre; not only should universities embrace digital OA, but also lead actions in cultivating diversity, equity, transparency and positive changes in society. Accordingly, this leads onto questions of whether we can evaluate the progress of OKIs and what are potential indicators for OKIs. As an exploratory study, this article reports on the collection and analysis of a list of potential indicators for OKIs. Data for these indicators are gathered for 43 Australian universities. The results show evidence of large disparities in characteristics such as Indigenous employment and gender equity, and a preference for repository-mediated OA across the Australian universities. These OKI indicators provide high-dimensional and complex signals that can be widely categorised into three groups of diversity, communication and coordination.

Keywords: open knowledge institutions; open access; diversity; principal components; hierarchical clustering.

1. Introduction

Demands on universities are changing, and universities need to change in order to meet these demands. Increasingly universities are interrogated about their effectiveness, impact, and accountability. The public wants to know how the taxpayer's money is used to drive positive changes in society. Students want to be able to evaluate the cost-effectiveness of the qualifications they pursue. However, the lack of clearly curated proxies for decision making,

together with facing resource limits, makes it difficult for universities to determine the best strategies to implement.

University rankings somewhat fill this gap by providing simple league tables of university performances. These rankings (and related metrics) have rapidly become dominant in influencing resource allocations, employment, student choices, management strategies, and beyond. They drive behavioural changes to various levels of the university ecosystem (Hazelkorn, 2008; 2009, Niles et al. 2020). However, the narrowly defined set of metrics used by these rankings are often low-dimensional and the information they provide are confined to very specific measures (Johnes, 2018; Selton et al., 2020). They are also often criticised for their methodological shortcomings and a number of unintended side effects (Kehm, 2014; Goglio, 2016).

More recently, there is a strong focus on making university research outputs more transparent and replicable. This is currently spearheaded by initiatives such as Plan S, which demands funder-supported research publications to be made open access (OA). While such movements for open science have gained momentum, there remain concerns about the broader commitment to openness in knowledge production and dissemination. There are already some signals of change in the way universities are evaluated, as evidenced by the Times Higher Education's Impact Rankings that assess performance against the United Nations' Sustainable Development Goals (SDGs), and the inclusion of OA and gender indicators in the CWTS Leiden Rankings. However, there is a need for a clearly structured framework that is able to capture a university's multidimensional efforts for achieving open knowledge goals.

Montgomery et al. (2020) describe one such framework in terms of *Open knowledge institutions* (OKIs). It advocates for universities to act with the principles of openness at their centre; not only should universities embrace digital OA, but also lead actions in cultivating diversity, equity, transparency and positive changes in society. The book also proposes an evaluation framework for OKIs, where potential indicators are categorised into three platforms of *diversity*, *communication* and *coordination*. This is combined with a theory of change that evolves through *aspiration*, *action* and *outcomes*.

Data and results reported in this article serve as a proof of concept for the OKI evaluation framework. As an exploratory study, we collect data for a number of OKI indicators for 43 Australian universities[1]. These include indicators related to OA, collaboration, output formats, physical and online accessibility, indigenous employment, gender equity, policy and infrastructure, and annual reports. We find a number of national anomalies such as large disparities in diversity of employment (which is negatively correlated to physical accessibility), and preferences for repository-mediated OA. Using robust statistical methods, we demonstrate that the signals provided by these indicators can be broadly categorised into the three platforms

[1] These include universities listed under Table A and Table B of the Higher Education Support Act 2003 (https://www.legislation.gov.au/Series/C2004A01234) and Avondale University College.

of diversity, communication and coordination. However, these signals are high-dimensional (i.e., diverse) with complex correlation structures, which also coincides with the theory of change described above.

The rest of the article is structured as follows. Section 2 introduces the data and indicators, with discussions on potential signals that these indicators may reveal. A brief description of statistical methodologies are also provided. Section 3 reports on the various data analysis. These begin with descriptive statistical analysis of the OKI indicators, followed by correlation analysis, principal component analysis (PCA), and cluster analysis, respectively. Section 4 discusses and summarises the main findings and implications thereof. Limitations of the study and conclusions are given in sections 5 and 6, respectively. Acknowledgements, a list of references, and a number of appendices follow.

2. Data and methodology

We have gathered information for a selected set of 26 OKI indicators, plus an indicator on university revenue, from a variety of data sources. While some of these were collected via semi-automated procedures, others were collected manually. We acknowledge that this list of indicators is not complete. However, it provides an important outlook for the challenges of data collection and integration, and the complexities in managing and interpreting a diverse set of potentially interconnected indicators, for OKIs. The dataset also provides a unique view of university performance at multiple levels and dimensions. Wherever possible, the data is focussed on the year 2017. The list of indicators examined, their data sources and collection processes are described in Table 1.

Table 1: The list of indicators, their description and data sources.

Indicator	Description	Data source and collection
oa_total	The proportion of outputs from a university that are made freely available online, either via the publishers or repositories.	Our data on OA were obtained through the Curtin Open Knowledge Initiative (COKI) data infrastructure. For each university, its outputs were collected from Web of Science, Scopus and Microsoft Academic. This is filtered down to outputs with Crossref DOIs. Subsequently, the OA status of each output is then obtained from Unpaywall. Readers are referred to Huang et al. (2020b) for details on this data collection process.
oa_gold	The university's proportion of outputs made freely available online via publishers under any OA license.	As above.
oa_bronze	The university's proportion of outputs made freely available online via publishers but with no clearly defined OA license.	As above.
oa_green	The university's proportion of outputs made freely available online via repositories,	As above.

	regardless of whether they are also available via the publishers.	
oa_green_only	The university's proportion of outputs made freely available online via repositories, but are not available via the publishers.	As above.
output_div	The coefficient of unalikeability (Kader & Perry, 2007) based on the types of outputs affiliated to the university. This measure ranges between 0 and 1, where a higher number is indicative of more diverse output types.	These output types are identified using Crossref's "type" field. Again, this is obtained via the COKI data infrastructure. The output types include "journal_articles", "book_sections", "authored_books", "edited_volumes", "reports", "datasets", "proceedings_article", and "other_outputs".
collab_total	The university's proportion of outputs co-authored with one or more other organisations.	Organisations are identified using unique identifiers from GRID (https://www.grid.ac/). This data is curated through the COKI data infrastructure, as for the OA data. Institutional links are drawn through co-authorships in Microsoft Academic. This is in addition to institutional search results for 1207 universities globally (including the top 1000 in the 2019 Times Higher Education World University Rankings) from the Web of Science and Scopus APIs.
collab_aus	The university's proportion of outputs co-authored with one or more other universities from the list of 43 Australian universities.	As above.
collab_other	The university's proportion of outputs co-authored with one or more organisations not from the list of 43 Australian universities.	As above.
collab_ind	The university's proportion of outputs co-authored with one or more industry partners.	This is obtained directly from the 2019 CWTS Leiden Ranking's indicator for industry collaboration (https://www.leidenranking.com/ranking/2019/list). This indicator covers data ranging from 2014 to 2017.
event_total	The university's proportion of outputs with at least one Crossref event.	This is determined by counting the number of outputs with existing events in the Crossref Events Data for each of the university's outputs and divided by the total number of outputs. Data is collected and curated via the COKI data infrastructure. See https://www.crossref.org/services/event-data/ for a list of data sources for events. Also see Appendix A for the justification of its use.
walk_score	This is an index of efficiency of the physical location of the university.	This is manually obtained from http://www.walkscore.com. This is included as a proxy for physical accessibility of the university.
web_score	This is a score assigned to the university's website based on the W3C Web Content Accessibility Guidelines (WCAG) 2.0 Level A and AA requirements.	This is obtained from the Functional Accessibility Evaluator 2.0 at https://fae.disability.illinois.edu/ on the status of the university's website at the last available day of 2017 through the Internet Archive (https://archive.org/web/). This is designated to

		indicate the university's effort to make their website more accessible.
indigenous	The university's proportion of indigenous staff (out of all staff).	Data obtained from the Australian Government's Department of Education, Skills and Employment website at https://docs.education.gov.au/node/46146
women_above_sl	The university's proportion of women, out of all academic positions above senior lecturer level.	Staff gender diversity data 2001-2018 downloaded as excel file from "uCube" http://highereducationstatistics.education.gov.au/Default.aspx using measures Staff Count, Current duties classification, Gender, Year, Institution. Staff count includes full-time and fractional full-time staff only.
women_sl	The university's proportion of women, out of all academic positions at senior lecturer level.	As above.
women_l	The university's proportion of women, out of all academic positions at lecturer level.	As above.
women_below_l	The university's proportion of women, out of all academic positions below lecturer level.	As above.
women_acad	The university's proportion of women, out of all academic positions.	As above.
women_non_acad	The university's proportion of women, out of all non-academic roles.	As above.
policy_lib	The university's score for library access policies.	Using Python scripts and checking manually on the university website for various characteristics. See Appendix B for more information.
policy_oa	The university's score for policies and support for OA publications and data.	As above.
policy_div	The university's score for policies in equity and diversity.	As above.
ann_rep_div	The proportion of phrases in the university's annual report that relate to diversity.	Annual reports (in PDF format) are collected from the university websites. Subsequently, a Python script is used to analyse word counts and number of occurrences of key phrases. See Appendix C for more detail.
ann_rep_comm	The proportion of phrases in the university's annual report that relate to communication.	As above.
ann_rep_coord	The proportion of phrases in the university's annual report that relate to coordination.	As above.
total_rev	The university's total revenue, recorded in thousands of Australian dollars.	Data collected manually from the Department of Education, Skills and Employment (https://www.education.gov.au/2008-2017-finance-publications-and-tables).

This diverse set of indicators is aimed at capturing various dimensions of OKIs. The inclusion of a number of OA indicators are intended to capture an OKI's level of commitment on the different routes of OA. The types of publication (e.g., journal articles, book chapters, etc.) is often associated with disciplinary practices and represents diverse ways in which the university engages with its surrounding community. Similarly, the different collaboration indicators signal the demographic and geographic reach of the university's research networks. A less traditional indicator is the proportion of outputs with Crossref events[2]. This is positioned as a signal of the university's practices for online engagement and visibility; it includes events such as social media mentions and references in Wikipedia.

The "walk_score" and "web_score" are intended to give some indication of a university's level of accessibility both physically (the former) and online (the latter). The "walk_score" takes into account walking distances and routes to nearby amenities, and pedestrian friendliness[3]. A score for each university's main webpage is obtained via the Functional Accessibility Evaluator. This tool evaluates websites based on the W3C Web Content Accessibility Guidelines 2.0 Level A and AA requirements. In essence this is aimed at measuring how closely a website satisfies a list of recommendations to make its content more accessible to people with disabilities and more usable by individuals with challenging abilities due to aging.

We also include a set of measures for diversity. In particular, we include the proportion of indigenous staff, and proportions of women within various categories of university positions[4]. These are signals of the university's efforts in being inclusive in its knowledge production. Wilson et al. (2020) provide a detailed description of the challenges in collecting and interpreting such data at both national and international scales.

Lastly, we have a number of indicators related to the university's policies and infrastructure surrounding a number of characteristics. These are intended to signal the university's efforts in coordinating various facets of an OKI internally and across external communities. The first three of these indicators relate to library access, OA, and diversity, respectively. In particular, a university's score for each of these policy indicators are related to a predefined set of characteristics surrounding policy statements, regulations and provision for support. Further details for these indicators are provided in Appendix B. We have also analysed each university's annual report for keywords and key phrases surrounding diversity, communication and coordination. The indicators "ann_rep_div", "ann_rep_comm" and "ann_rep_coord" are constructed as the number of times keywords or key phrases (from the predefined

[2] A Crossref event is defined as an instance of mention or reference of a research output recorded over the web via the Crossref Event Data. This can come from a number of sources such as Twitter, Wikipedia, Newsfeed, etc.
[3] https://www.walkscore.com/methodology.shtml
[4] As per Department of Education, Skills and Employment.
https://www.education.gov.au/higher-education-statistics

corresponding lists) appears in the document, divided by the total number of words in the document.

We supplement the list of OKI indicators above with the total revenue received by each university. This is aimed at serving as a benchmark for size and prestige. At the same time, we are interested in how total revenue may affect the OKI indicators.

It should be noted that our data does contain missing values. In particular, there are missing values in “collab_ind”, “web_score”, each of the gender indicators, each of the annual report indicators and “total_rev”. Our subsequent analysis takes these missing values into account where applicable. These are differentiated from true zeros, which also exist in the data.

We utilise several statistical techniques to analyse this set of data. First, we focus on statistical descriptions that allow better comparisons across the different indicators. This is followed by an in depth analysis of Spearman’s rank correlation between the indicators, assessing potential monotonic relationships between pairs of indicators. Robust PCA approaches are used to explore how the total variance across the OKI indicators can be best described by a smaller set of orthogonal principal components (PCs), and how the indicators relate to these PCs. Finally, cluster analysis performed on the data reveals clusters of universities using the OKI indicators as the clustering criterion.

3. Data analysis

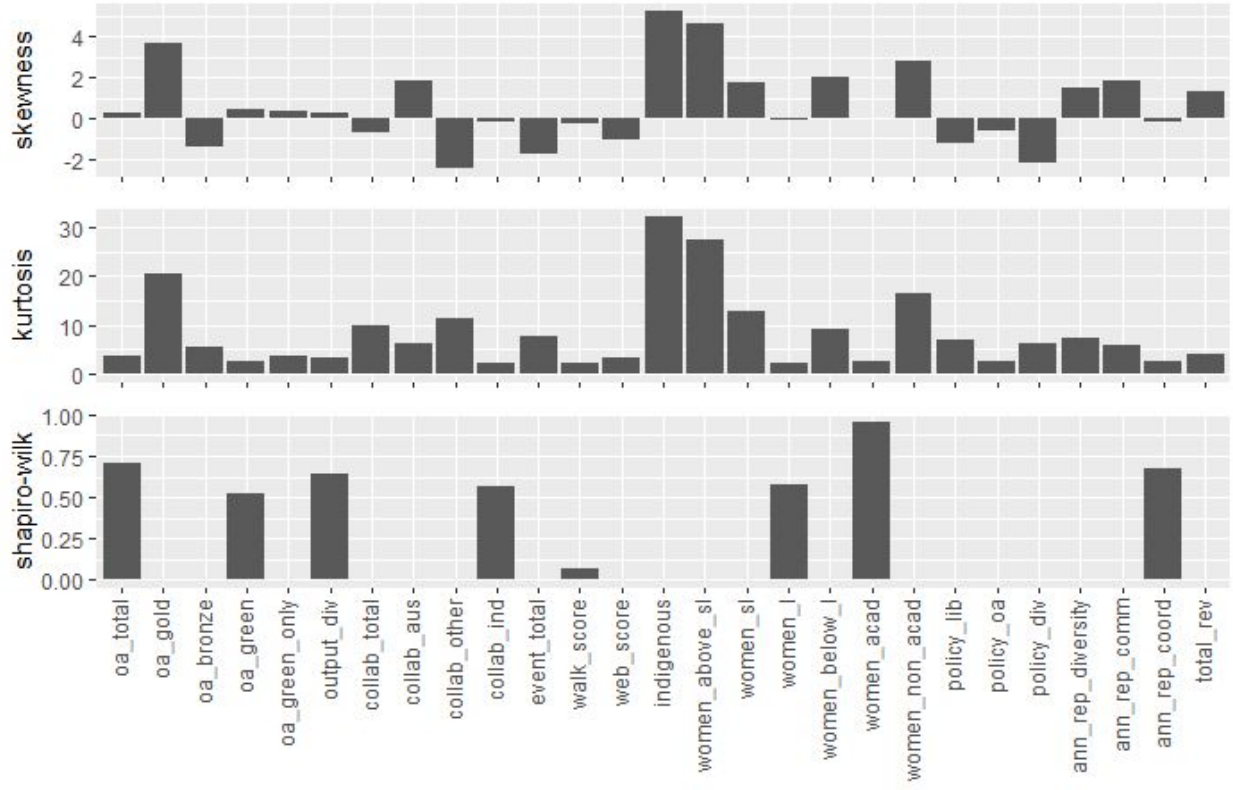
3.1 Descriptive analysis

Our analysis starts with some individual descriptive statistics, and seeks to get insight into patterns emerging from each indicator and comparisons thereof. A number of summary statistics are recorded in table D1 of Appendix D for reference. Here we focus on several statistical descriptions that are more comparative across the different measurements. Figure 1 below presents each indicator’s level of skewness, level of kurtosis, and its p-value resulting from the Shapiro-Wilk normality test. Many of the indicators exhibit substantial skewness (deviation from 0), and are leptokurtic (kurtosis value greater than the normal distribution, i.e., 3). These characteristics are consistent with the low p-values obtained for the corresponding Shapiro-Wilk test, indicating that many of the indicators are highly unlikely to be normal in distribution.

Proportion of indigenous staff (“indigenous”) displays the highest levels in both skewness and kurtosis. It is positively skewed due to two universities having 16.13% and 4.4% indigenous staff, respectively, compared to all other universities sitting in the 0% to 2.4% range. Other indicators with highest levels of skewness and kurtosis include “oa_gold”, ”collab_other”, ”women_above_sl”, and “women_non_acad”. For each of these indicators, the data contains

extreme[5] observations. These extreme data points may have significant influences on the measures presented in Figure 1.

Figure 1: Skewness, kurtosis and Shapiro-Wilk normality test p-value for indicators.



For an alternative view of each indicator's distribution, we construct their respective histograms. We use the Freedman-Diaconis rule for binwidths (hence, the corresponding number of bins) which is less sensitive to extreme observation than the standard Sturges' formula for number of bins. This is applied to all indicators other than "policy_lib", "policy_oa" and "policy_div", where the values of the original observations are used as bins. These histograms are displayed in Figure 2.

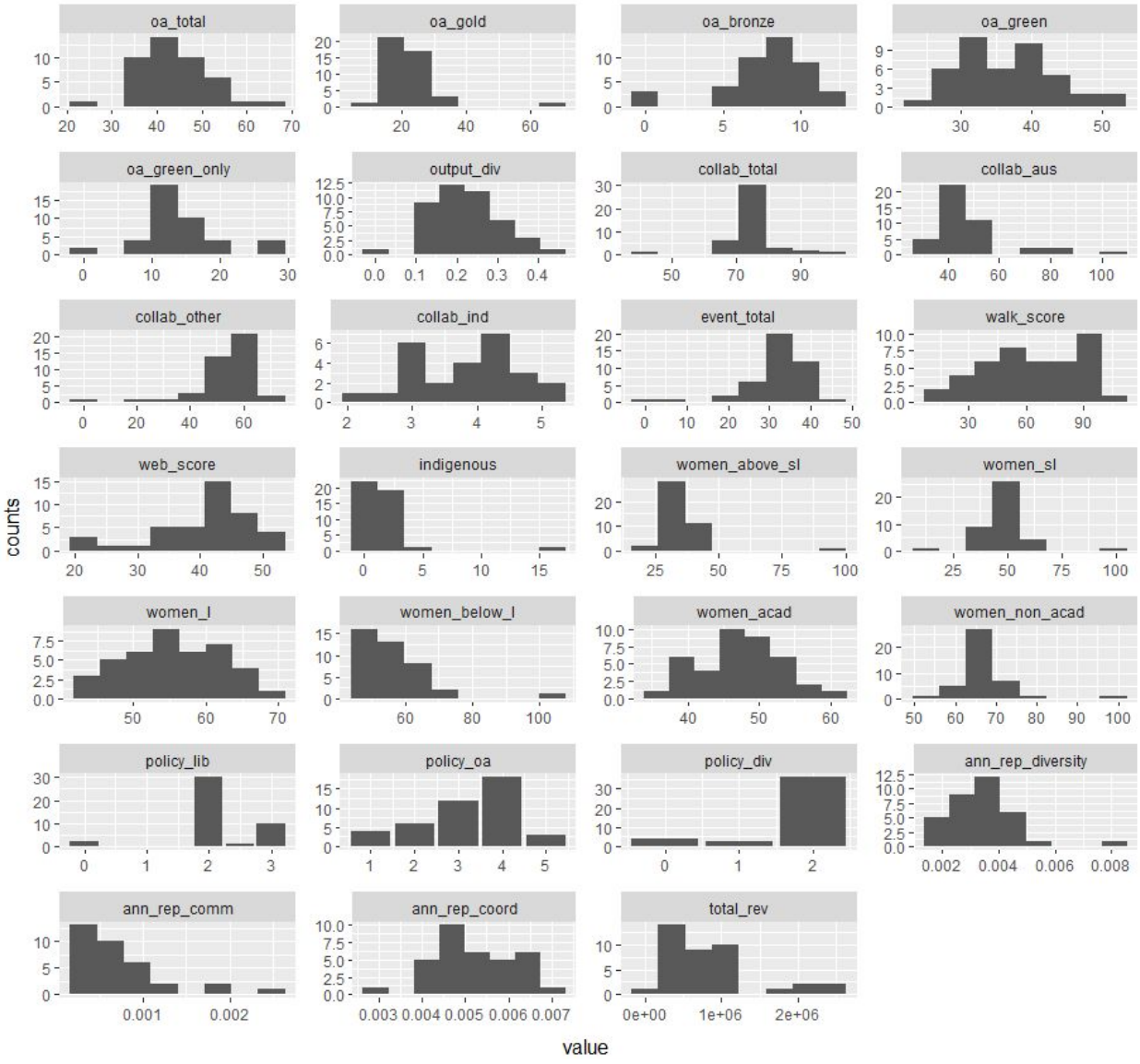
As mentioned, many of the indicators exhibit extreme observations. However, even with these extreme observations ignored[6], many indicators are still characterised by substantial skewness (i.e., asymmetry) and high levels of kurtosis (i.e., fat tails).

To better understand each indicator, it is worth discussing some of these extreme observations. For example, the single extremely large observation for "oa_gold" corresponds to a significantly smaller institution in output size where 2 out of 3 publications were made OA via the publisher.

[5] These are not outliers due to error and cannot be simply dropped.
[6] This may be more clearly presented in the boxplots of normalised observations. See Figure D1 of Appendix D.

The same university is responsible for the largest observations in “collab_total”, “collab_aus”, “indigenous”, “women_above_sl”, “women_sl”, and the lowest values for “collab_other”, “collab_ind”. Similarly, three of the smallest universities have 0% for “oa_bronze” and two of the smallest universities also contribute to the two lowest values in “event_total”. However, some extreme observations also result from other universities. For example, a medium to large sized university resulted in the largest value for “ann_rep_diversity”.

Figure 2: Histograms of indicators.

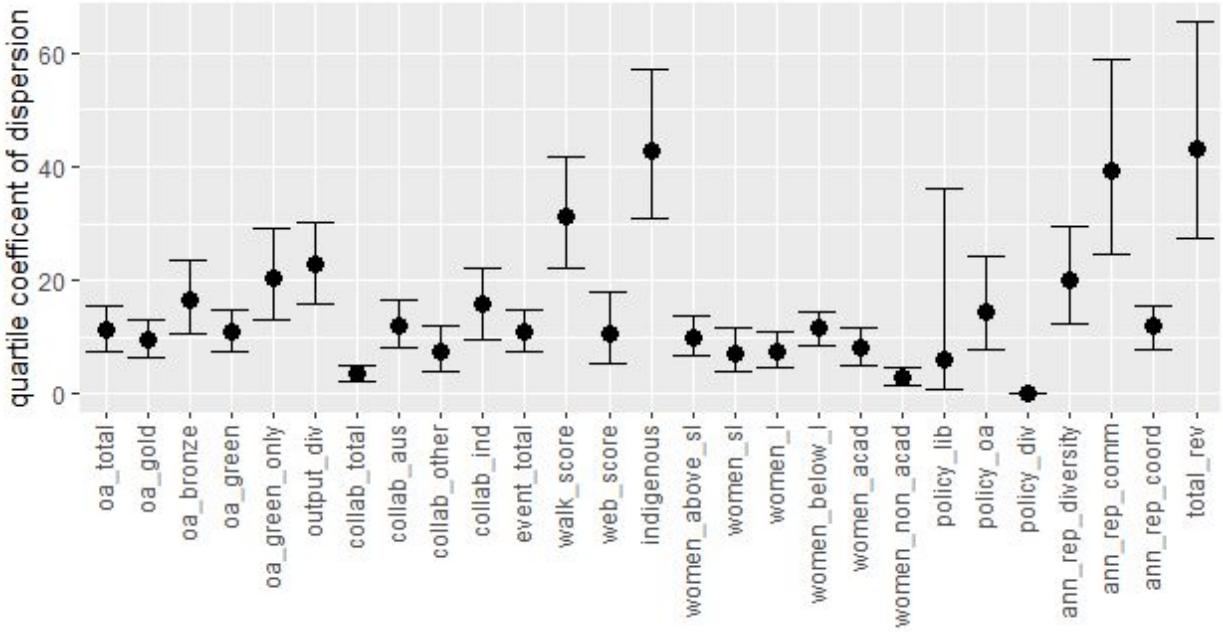


All 43 universities, apart from one, have a higher “oa_green” proportion than “oa_gold”, which is expected as most gold OA publications are also archived by repositories. In contrast, 34 of the universities have higher values in “oa_gold” than in “oa_green_only”. This is in agreement for findings on a more global scale for 2017 (Piwowar et a., 2018; Montgomery et al., 2020).

Australian universities also have a low uptake on the Bronze OA route. 29 out of the 43 universities actually have their “collab_other” proportions higher than the corresponding “collab_aus” proportions. This indicates that these universities have more instances of collaboration with research organisations outside the 43 Australian universities than among themselves. The indicator “walk_score” aims to signal the ease of physical accessibility of a university. As expected, universities located at city centres resulted in high scores, while lower scores mostly correspond to universities in regional areas (albeit with a few exceptions).

The existence of extreme observations, together with highly diverse measurement units across the indicators, can easily distort standard measures of spread, such as variance and coefficient of variation. Hence, we supplement our analysis by calculating the quartile coefficient of dispersion (QCD) for each indicator. QCD is a unitless measure of dispersion that is more robust against extreme observations. This allows us to make an overall comparison of the relative dispersion across the indicators, and also study the level of disparity within each indicator. The results are displayed in Figure 3, together with confidence intervals[7] for the QCD value in each case.

Figure 3: Quartile coefficient of dispersion (with confidence interval) for indicators.



Immediately we observe that “total_rev” and “indigenous” have the highest values for QCD, an indication of their high inequalities across the universities. The large inequality of revenue is highly correlated with output size (i.e., 0.97 in Spearman’s coefficient of rank correlation). This is immediately followed by the high level of disparity in the proportion of indigenous staff (even with the effects of the extreme observations minimised by QCD). “ann_rep_comm” also has a

[7] Bonett’s confidence interval, as appropriate for the sample sizes here (Bonett, 2006; Altunkaynak & Gamgam, 2019).

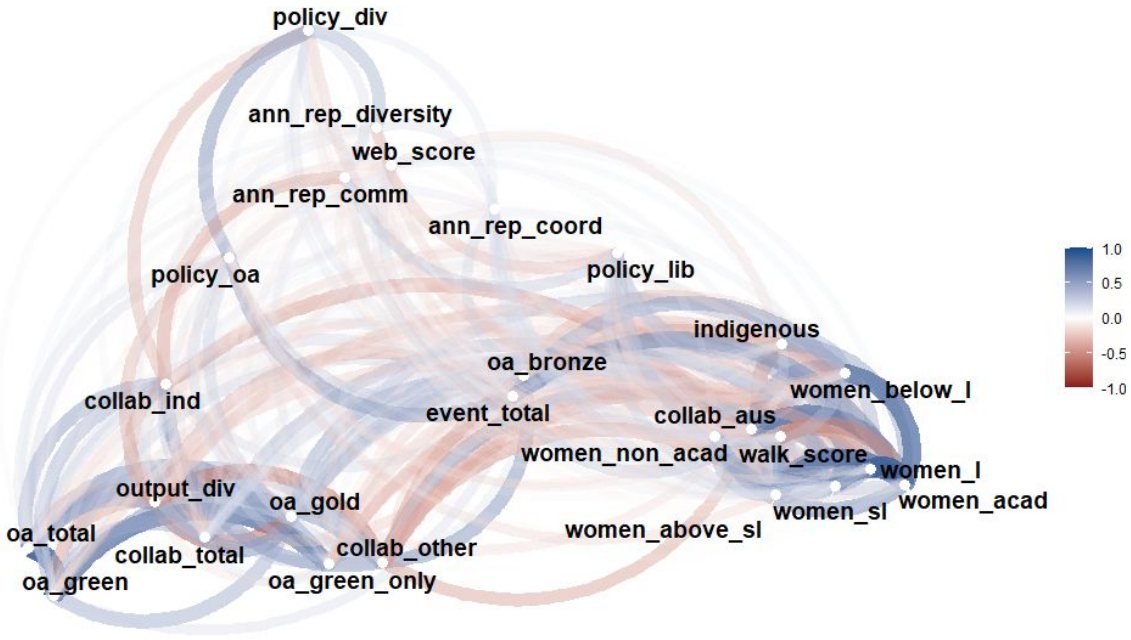
high QCD value, signalling the large differences across university annual reports in terms of the proportions of phrases or keywords related to communication. The “walk_score” indicator has the next highest QCD, which is likely attributed to the large differences in physical accessibility due to locations of the universities.

The proportion of collaborative publications (“collab_total”) and the proportion of non-academic women staff (“women_non_acad”) have the lowest QCD value (ignoring the trivial case of “policy_div”). This is a result of high concentration of values around the central part of their respective distributions (around 74% and 66% respectively). Overall, there is a high degree of differences in disparity across the indicators.

3.2 Correlation analysis

In this section, we explore the potential relationships between the OKI indicators. We calculate the Spearman’s rank correlation coefficient between all pairs of indicators[8]. The overall result is summarised into a network plot presented in Figure 4. In the Figure, indicators that are more highly correlated appear closer together, and are joined by links with darker shades. Blue links indicate positive correlations, while red links represent the negative correlations.

Figure 4: A network plot of Spearman’s rank correlation between OKI indicators.



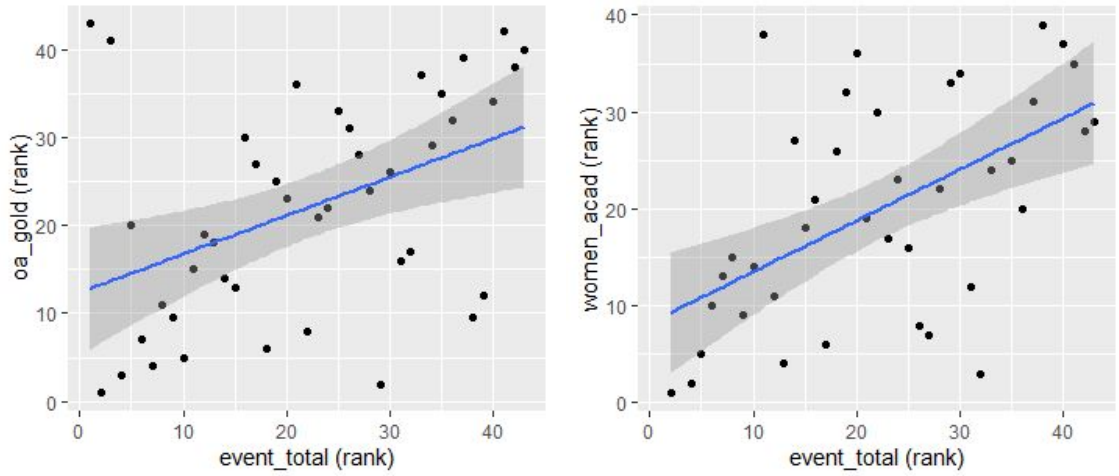
In Montgomery et al. (2020), an evaluation framework for OKIs is proposed by characterising indicators into three platforms: *diversity*, *communication* and *coordination*. This is combined with

[8] Missing values are left out and only pairwise-complete observations are included for each pairwise calculation.

a theory of change that evolves through the three stages of *aspiration*, *action* and *outcomes*. It also noted that the indicators may become increasingly more difficult to characterise into the three platforms as we move through the three stages of change. Figure 4 represents a practical example of the above. In the bottom right, we see a cluster of indicators mainly related to diversity (i.e., gender and indigenous proportions). The bottom left section of the Figure encompasses a group of indicators related to communication, such as OA and collaboration. And, perhaps more dispersed, is a set of indicators for coordinations (e.g., policies) gathered at the top of Figure 4. Note that "total_rev" is not included in Figure 4, as it is not directly considered as an OKI indicator. But we will take a closer look at its correlations with the OKI indicators later. It should be noted that the proximities of the points in Figure 4 are determined by multiple clustering. This means where a point lies is relative to its magnitudes of correlations with all other points. Hence, direct comparisons between correlations of different pairs of indicators must be made with caution.

A number of indicators appear less clearly defined in terms of which platforms they relate to, such as "event_total" and "oa_bronze". This does not necessarily imply that they are not correlated with other OKI indicators. Rather, they may be similarly correlated (in magnitude) to indicators from multiple platforms making them less distinctive for classification. For example, "event_total" has similar magnitudes of correlation with both the OA indicators and the diversity indicators (see Figure 5 for examples). If we consider "event_total" as a measure of action or outcome (rather than aspiration), then its indistinctive positioning conforms again with the theory of change within the OKI evaluation framework.

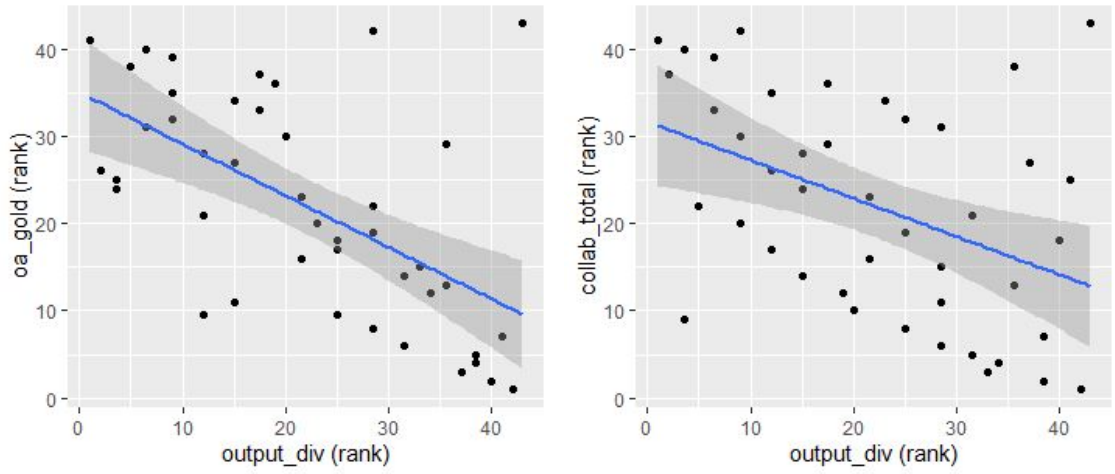
Figure 5: Scatterplots between ranks in "event_total" versus ranks in "oa_gold" and "women_acad", respectively.



Another interesting observation is that "output_div" being closely located to many of the OA indicators and the collaboration indicators. These correlations are in the negative direction throughout. Two examples are given in Figure 6. We should note here that most of the bibliographic data systems (including ours) are better at capturing accurate information on

journal articles than other output types. Also, given most universities have journal articles as the primary output format, the “output_div” indicator is mainly influenced by the inclusion of other output types, such as book chapters, conference proceedings, datasets, etc. These other output formats are also more likely to be recorded as non-OA and are affiliated to smaller numbers of authors.

Figure 6: Scatterplots between ranks in “output_div” versus ranks in “oa_gold” and “collab_total”, respectively.



The indicators “collab_aus” and “walk_score” appear out of place at the first glance. They seem more highly correlated to the gender and indigenous indicators, and less correlated to other indicators relating to collaboration and access. However, a deeper exploration reveals some interesting relations. The “walk_score” is negatively correlated with the diversity indicators, while “collab_aus” is correlated positively with the same indicators. Examples of these for “women_acad” are given in Figure 7. Note that “walk_score” can also be seen (in general) as a benchmark for the locations of universities. This implies universities in more regional areas (which also tend to have lower total revenue and are smaller in size) have proportionally more women and indigenous staff. These universities also produce higher proportions of outputs co-authored with others in the list of 43 universities (hence a negative rank correlation between “walk_score” and “collab_aus”). In contrast, “walk_score” is positively correlated with the other collaboration indicators, albeit at relatively smaller magnitudes.

We also note the generally low correlation levels between OA indicators and the gender indicators. In comparison, Olejniczak & Wilson (2020) found that (for a sample of US universities) there is a slight bias towards male authors in terms of OA publications, and this bias increases if job security and level of resources are taken into account. Our data present little to no evidence for this bias in the Australian context, although further study is required.

Pairs of indicators with the highest positive values of rank correlation are “oa_total” against “oa_green” (0.89), and “women_l” against “women_acad” (0.86). In contrast, pairs with the

highest negative correlation coefficients are “oa_gold” against “output_div” (-0.59), “output_div” against “collab_other” (-0.57), and “indigenous” against “walk_score” (-0.58). We have already had discussions related to pairs with highest negative correlations earlier. So we will now focus on pairs with the highest positive correlations. Figure 8 displays the scatterplots of the two pairs with the highest positive correlations.

Figure 7: Scatterplots between ranks in “women_acad” versus ranks in “walk_score” and “collab_aus”, respectively.

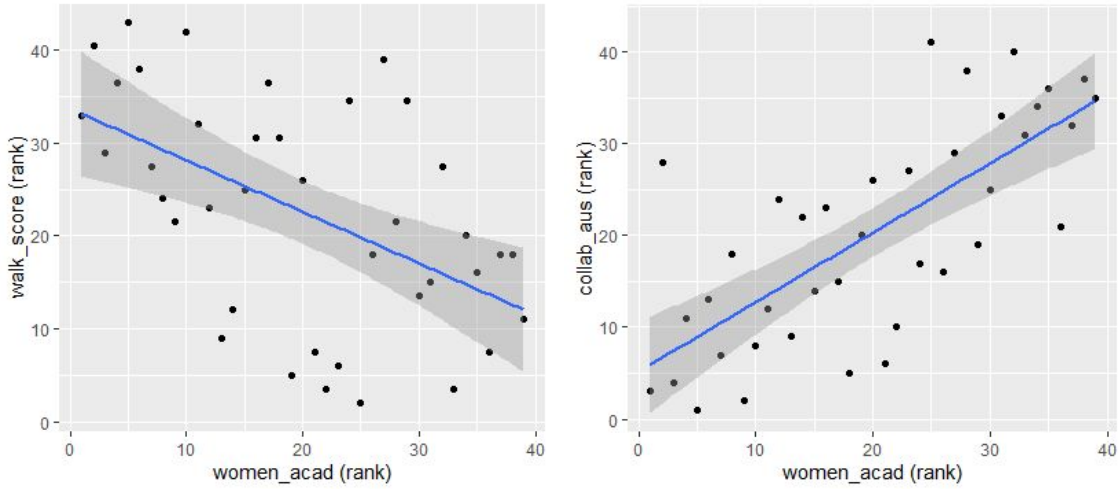
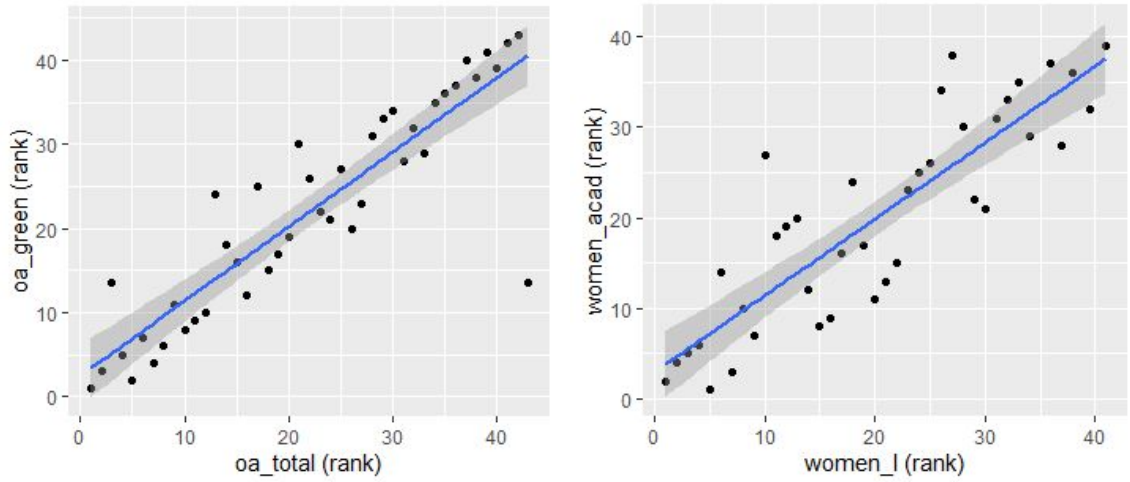


Figure 8: Scatterplots between ranks in “oa_total” versus “oa_green” and “women_l” versus “women_acad”.



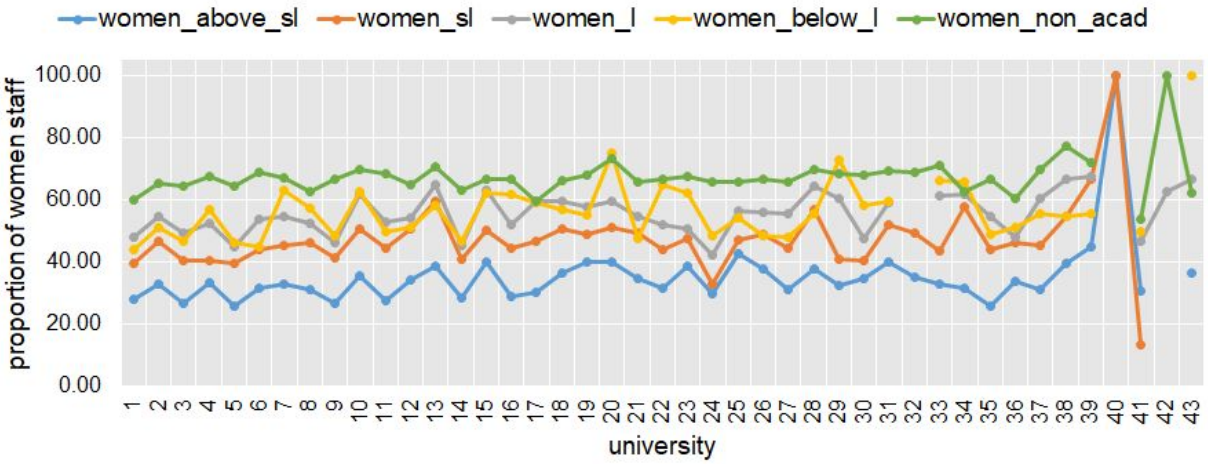
Recall that we have noted almost all universities in our study have a higher proportion of green OA than gold OA. Green OA (i.e., repository-mediated OA) is seen as the more cost effective route compared to gold OA (i.e., publisher-mediated OA) due to the high cost of article processing charges (APCs). Many outputs included as gold OA are also available through green OA, albeit usually of earlier manuscript versions. Huang et al. (2020b) provided a parallel view

at a more global scale, depicting a much more diverse set of OA paths. In comparison, the pattern emerging in Australia is more uniform. That is, there is a relatively stronger overall focus on green OA than gold OA (note that the rank correlation coefficient is 0.55 between “oa_total” and “oa_gold” and is 0.89 between “oa_total” and “oa_green”).

The strong correlation between “women_l” and “women_acad” requires a deeper exploration. Figure 9 provides an overview of the proportion of women at various levels of academic positions within all 43 universities. Evidence shows there are higher proportions of women staff at lower academic positions consistently across all institutions. For example, 40 out of the 43 universities have a higher proportion of women in positions below lecturer level than proportion of women above senior lecturer level. These highlight the increase in gender inequality as we move up in the ranks of academic positions (Winslow & Davis, 2016; Baker, 2016). There is also a higher proportion of women in non-academic roles, where 42 out of the 43 universities have a higher value for “women_non_acad” than for “women_acad”. It is also higher than all other gender proportions for almost all universities (as depicted in Figure 9). Note that we also observed earlier (Figure 3) that “women_non_acad” has one of the lowest QCD values among all indicators.

Together with a much larger number of women employed at the lecturer level in academic roles, the above are factors that contribute to the high correlation observed between “women_l” and “women_acad”. These findings are consistent with existing literature which finds progress for gender diversity to remain at more junior and non-academic positions within universities. This is despite numerous reported diversity policies and action plans by universities (Khan et al., 2019; Marini & Meschitti, 2018; Subbaye & Vithal, 2017; Winslow & Davis, 2016; Baker, 2016)

Figure 9: Proportion of women at various levels of academic positions for all 43 Australian universities (anonymously labelled 1 to 43).

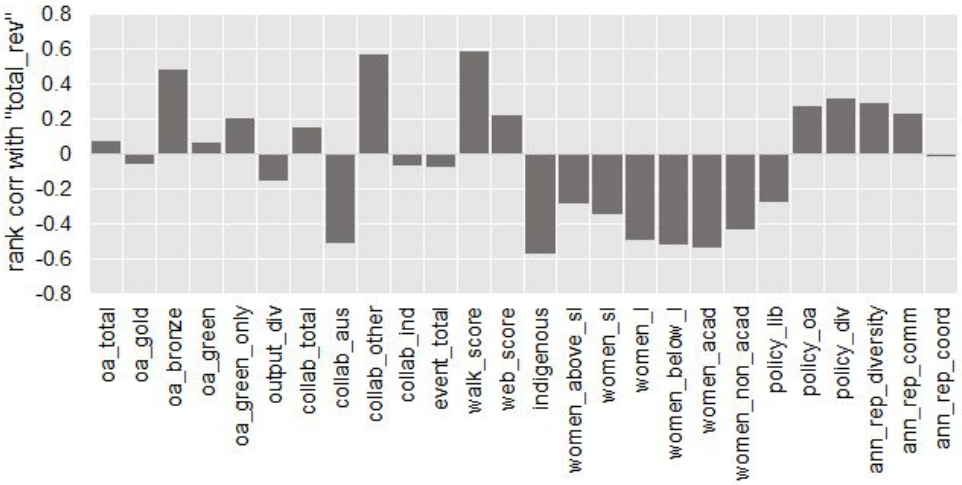


Lastly, we look at the potential influences of the university’s revenue on the OKI indicators. Figure 10 displays the rank correlation coefficient values between “total_rev” and each of the

other indicators. The three-way correlation between “total_rev”, “walk_score” and “indigenous” is consistent with our comment earlier regarding the sizes and locations of universities. That is, smaller, regional, and less wealthy universities are correlated with higher proportions of indigenous staff. The negative correlation between “total_rev” and all the gender indicators also displays a similar pattern. At least in theory, this potentially implies indigenous and women academics may need to relocate to seek potential promotion opportunities. Both of these trends re-emphasise barriers to progress in hiring practices and achieving equity and parity (especially at high ranking institutions). However, the levels of Indigenous employment may be driven by local and regional demographics, and recognising such differences may be vitally important for actioning policies and support for change. It is also worth noting that almost all universities in our data have existing policies on employment equity and diversity (as recorded by the “policy_div” indicator). Evidently, some differences in outcomes exist across universities, but the levels of actioning (upon existing policies) remain difficult to quantify and require further exploration.

The correlation of “total_rev” versus “collab_aus” and “collab_other”, respectively, are high but in opposite directions. Consistent with our earlier discussion, universities with lower total revenue appear to have more proportions of output co-authored with other universities in our list of 43 Australian universities. In contrast, the more wealthy universities seem to have higher collaboration proportions outside these 43 universities, including international universities and research organisations. This may be attributed to the fact that international collaboration enhances an institution’s reputation, impact, and ability to attract research and development investments, and research talents through both researchers and students (Australian Academy of Humanities, 2015; Glänzel, 2001; Glänzel & de Lange, 2002). These in turn influence the university’s position in international rankings. The size of “total_rev” seems to have little correlation with “collab_ind”, though this indicator is derived from an external source directly and should be interpreted with caution.

Figure 10: Spearman’s rank correlation coefficient of “total_rev” versus each of the OKI indicators.



The generally low correlations between “total_rev” and the OA indicators (except for “oa_bronze”) is an interesting outcome. They are a potential indication that higher revenues at Australian universities do not necessarily translate to higher proportions of OA outputs. This is in contrast to Siler et al. (2018) which suggested gold OA publishing to be correlated to levels of funding and university ranks, though this is restricted to the field of Global Health research. Our findings do however conform to the fact that only 3 universities in our data have signalled a provision of funding for OA publishing. This, and the fact that all but one university have an OA repository, may explain the slight increase in correlation against “oa_green_only”. We should also note that the OA indicators seem to have only low to moderate correlations with “oa_policy”. The moderate correlation between “oa_bronze” and “total_rev” poses an interesting case, where we find more wealthy universities to have higher proportions of Bronze OA publications. However, these proportions remain generally low for all universities in the study.

3.3 Principal components analysis

In this section, we apply PCA on the OKI indicators. This is aimed at providing insight into how information is attributed across the different indicators, and how these indicators relate to a few principle components (PCs). Due to the existence of extreme observations, missing data and vastly different measurement scales, with no existing knowledge of a plausible re-scaling method, we propose a two stage process for PCA.

Firstly, being constrained by the size of data, we propose imputing the missing values using an iterative PCA algorithm[9]. This procedure uses the mean of each indicator as initial values for the missing data. Subsequently, a standard PCA is applied and a selected number of PCs are used to re-estimate the missing values. The process is repeated iteratively until the imputed values convergence.

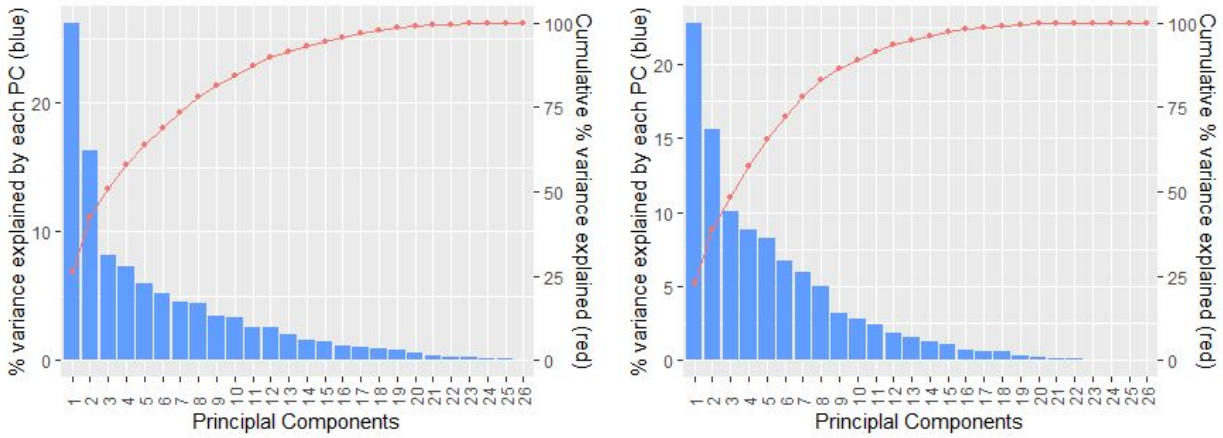
Once the imputed data is obtained, we proceed with robust PCA procedures that cater for extreme observations. Common approaches for this purpose include the use of robust covariance (or correlation) matrices and projection pursuit. To confirm the robustness of our result, we implement two different methods. The first is the use of the Spearman’s rank correlation matrix[10] in the classical PCA procedure. Alternatively, we apply the ROBPCA procedure which combines robust covariance matrix estimation with projection pursuit (Hubert et al., 2005). We discuss some of the major findings below, with supplementary results provided in Appendix E.

Figure 11 shows the percentages of variance explained by each of the PCs derived from each of the two PCA approaches, respectively. The corresponding cumulative percentages of

[9] This is implemented by using the R package `missMDA`. The function `estim_ncpPCA` is run to estimate the number of PCs to be used for imputation. This is followed by the iterative PCA process run by the `imputePCA` function.
[10] Equivalent to applying standard PCA to ranks within each indicator.

variance is recorded by the red line. It is observed that we need at least 8 PCs to attain a variance coverage of approximately 80% for the Spearman PCA. For the ROBPCA approach, we need 7 PCs to attain a similar level of variance coverage. Analogous results are observed using the Kaiser criterion on eigenvalues to determine the number of PCs to retain (see Figure E1 in Appendix E).

Figure 11: Percentages of variance explained by PCs (with cumulative variance) from PCA with Spearman's rank correlation (left) and ROBPCA (right).



The low proportions of variance explained by the individual PCs and the high number of PCs needed to attain a significant coverage of the cumulative variance indicate that the set of OKI indicators provides diverse information where the overall variance is spread across multiple directions. However, keeping a high number of PCs also makes the interpretations of these PCs more difficult, as the later PCs show less distinctive groupings of loadings by the original indicators (see Tables E1 and E2 for the loadings on the first 8 and 7 PCs, respectively for each PCA approach). The first two PCs display more resemblance of the OKI evaluation framework described earlier, with one of the first two PCs correlated largely to diversity indicators while the other with the communication indicators. To a lesser extent, there may be a third PC that relates to coordination. As graphical illustrations of these, the correlation circle plots against the first two PCs are provided in Figures 12 and 13.

In the correlation circles, all 26 OKI indicators are projected onto the first two PCs from each PCA approach. Angles between arrowed lines represent correlations between indicators in this plane (with 90 degrees indicating zero correlation and 180 degrees indicating perfect negative correlation). Lengths of the arrowed lines are indicative of how well they are represented in this two-dimensional space (or their levels of contribution to these PCs). In Figure 12, many of the diversity indicators are pointing to the right along the first PC, while many of the communication indicators are pointing in the direction of the second PC. Similar pattern is observed in Figure 13. Many of the correlations represented in these correlation circles are also consistent with observations made in Figure 4 earlier. Most of the coordination indicators have shorter arrowed

lines, indicating they are not well-represented in this space, and are likely to be more representative by other PCs.

Figure 12: Correlation circle against first 2 PCs using Spearman PCA.

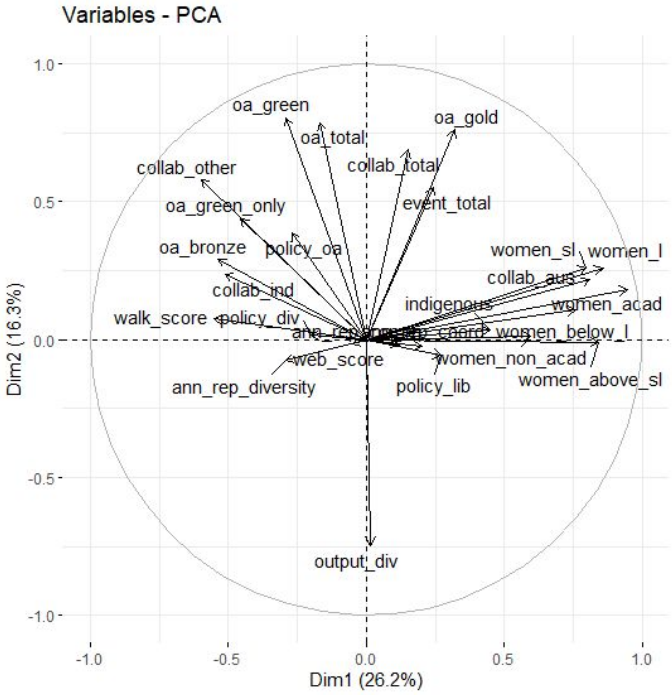
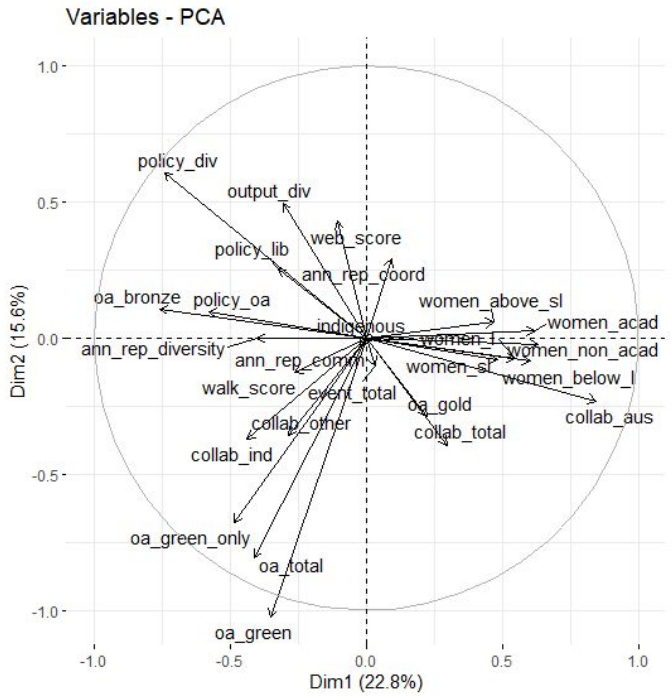


Figure 13: Correlation circle against first 2 PCs using ROBPCA.



The observations made above lead to a focus on the first three PCs (a third PC included to try capture variances in the coordination indicators). Table E3 in Appendix E lists the loadings by each of the OKI indicators after rotation[11] against the first three PCs. For a simplistic overview that aligns with the OKI evaluation framework, we summarise the results into what proportions of each PC's variance[12] are loaded by each of the three groups of indicators[13]. The groupings are decided depending on observations made in Figure 4, and in conjunction to the Spearman's rank correlation matrix and PCA loadings. Three of the indicators, "event_total", "oa_bronze" and "web_score", with non-distinctive grouping are listed separately. These are presented in Table 2.

The results in Table 2 re-affirms our discussion earlier regarding which groups of variables provide the most loadings on the first two PCs. In addition, we observe the significant level of loading of the coordination indicators on a third PC. The indicator "event_total" loads on the coordination PC in the Spearman PCA approach, but loads on the communication PC under the ROBPCA. This is again indicative of the complex relationships between "event_total" and other indicators. On the other hand, "collab_ind" and "oa_bronze" seem to have little influence on the first 3 PCs and the inclusion of more PCs is potentially needed to capture this information.

Table 2: Proportion of PCs' variances loaded by groups of OKI indicators.

	Spearman PCA			**ROBPCA**		
Platforms	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Diversity	**78.2%**	2.7%	15.5%	**66.3%**	3.6%	12.1%
Communication	10.1%	**84.3%**	7.9%	15.3%	**77.6%**	5.7%
Coordination	3.6%	3.4%	**61.6%**	3.2%	10.8%	**79.1%**
"event_total"	2.3%	4.6%	13.6%	9.9%	2.5%	2.9%
"oa_bronze"	2.9%	2.9%	1.1%	0.4%	2.6%	0.2%
"collab_ind"	2.9%	2.1%	0.3%	4.9%	2.9%	0.0%

As a summary of the PCA results we note the following. A relatively high number of PCs is required to capture the multidimensional variance provided by the OKI indicators. This is in contrast to many parallel analyses on popular university rankings where only 2 or 3 PCs are needed to explain a large portion of variance in their data (Dehon et al., 2010; Docamp, 2011l Selten et al., 2020). Furthermore, we show that the major contributors on the first three PCs (after rotation) can be grouped into three sets of OKI indicators of diversity, communication and

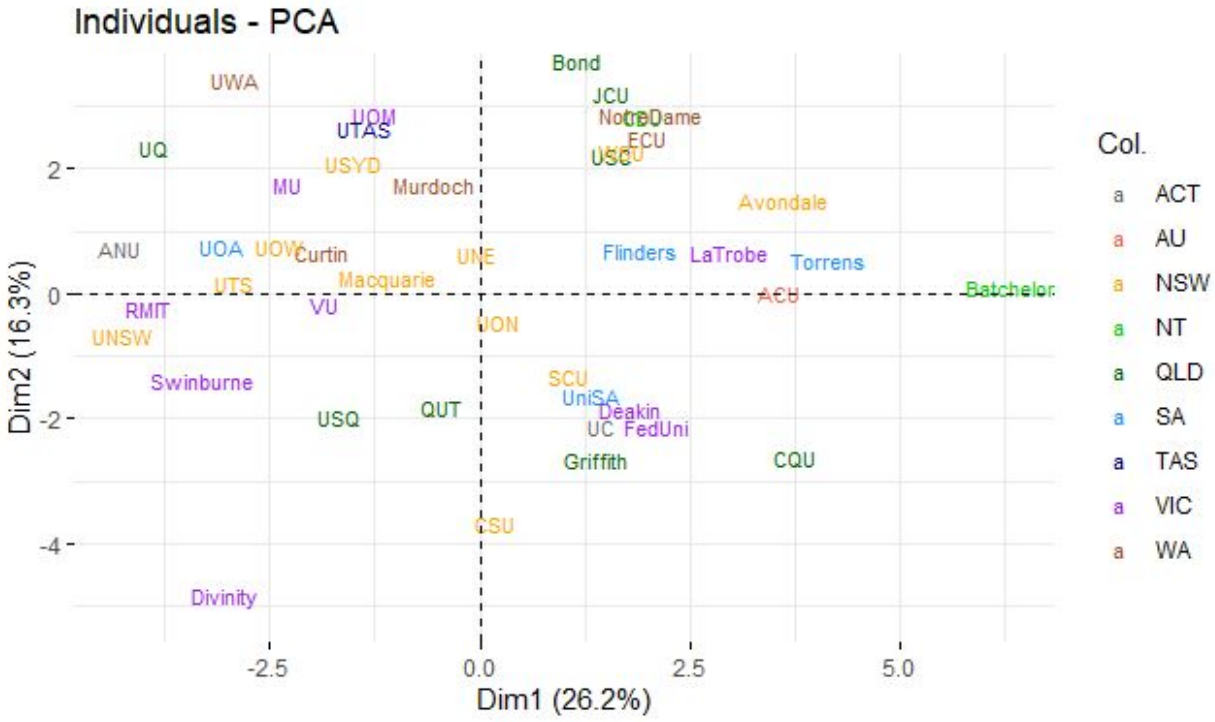
[11] Rotations are performed using the `varimax` function in `R`.
[12] This can be calculated by sums of squares of standardised loadings of the selected indicators.
[13] Diversity indicators: "indigenous", all gender indicators, "collab_aus" and "walk_score"; communication indicators: "oa_total", "oa_gold", "oa_green", "oa_green_only", "output_div", "collab_total" and "collab_other"; coordination indicators: "web_score" and policy and annual report indicators. Others: "event_total", "oa_bronze", "collab_ind".

coordination. This is a finding that is consistent with the OKI evaluation framework suggested by Montgomery et al. (2020).

3.4 Cluster analysis

In this section we are interested in exploring whether there are specific groupings of universities that can be defined by the OKI indicators. This is important in performing likewise comparisons and for identifying different paths of OKIs. As an immediate follow up from the PCA analyses, we are able to construct individual component plots of universities mapped onto any pair of PCs. In these plots, universities having similar profiles (or scores) in terms of the selected PCs will be displayed closer together. Figure 14 displays the individual component plot for the first two PCs from the Spearman PCA, with universities colour-coded by state or territory (as per main campus location). No immediate pattern arises in terms of universities from a common state or territory as each group seems to be randomly scattered.

Figure 14: Individual component plot for first two PCs from Spearman PCA, with universities coloured by state[14].

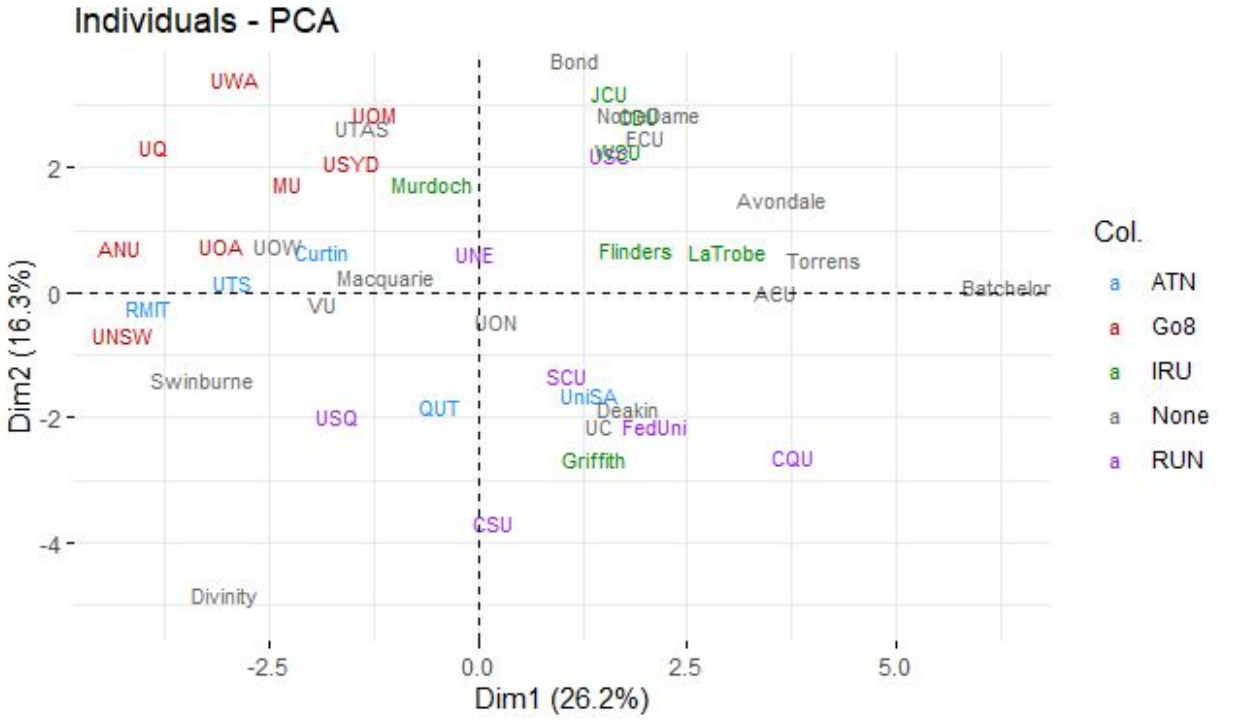


Alternatively, we assign a colour to each university by their affiliation to existing Australian university networks in the same plot. This is presented in Figure 15. The standout group is the Go8 where all 8 member universities lie towards the top-left of the plot. This is an indication that

[14] Australian Capital Territory (ACT), Multi-state (AU), New South Wales (NSW), Northern Territory (NT), Queensland (QLD), South Australia (SA), Tasmania (TAS), Victoria (VIC), Western Australia (WA).

these universities are quite similar in terms of their performance in the first PC (diversity) and the second PC (communication). Their overall performance leans toward the top half in communication, but tends toward the opposite direction for diversity. Parallel outputs from the ROBPCA are included in Appendix E, in Figures E2 and E3. They display similar findings albeit with more extreme positioning of some universities due to size differences being taken into account.

Figure 15: Individual component plot for first two PCs from Spearman PCA, with universities coloured by university network[15].



We next consider cluster analysis of universities using the full set of OKI indicators and the respective ranks. Columns are standardised[16] to cater for the different measurement levels. Hierarchical clustering is implemented by using the Manhattan distance[17] to construct the dissimilarity matrix between universities, and the complete-linkage criterion is used to select similar clusters[18]. These selections are made based on their robustness against extreme values.

[15] Australian Technology Network (ATN), Group of Eight (Go8), Innovative Research Universities (IRU), Regional Universities Network (RUN), and universities unaffiliated to any network groupings are labelled as "None".

[16] Columns are standardised by subtracting the column mean and dividing by the column's mean absolute deviation.

[17] Missing values are ignored as only pairwise-complete data are used.

[18] These are implemented using a number of functions and packages in `R`. The `daisy` function from `cluster` is used to calculate the Manhattan distance matrix. `hclust` is used for the cluster analysis and subsequently converted to a dendrogram object for plotting using the `dendextend` package.

Dendrograms of the clustering results are presented in Figures 16 and 17. The former is derived using ranks as input, while the latter used the original observations. In both figures, the university labels are colour-coded by university network affiliations as before. Corresponding figures colour-coded by states and territories are given in Appendix E. In comparison, the university networks reveal more synchronised groups in comparison to locations. The most prevalent case is that of the Go8 universities. In both figures, these universities appear to be closely clustered. This is a potential indication of the synergies across universities in common affiliated networks.

A further interesting observation is related to those universities that largely remain in singular or very small clusters. These extreme cases are less obvious in Figure 16 given the use of ranks removes the size effects. However, both figures consistently show that many of the last few universities to be added to clusters are small, private or specialist universities. Intuitively this makes sense given that such universities may have less resources, missions that deviate from traditional universities, and practices that need to be aligned to these.

Figure 16: Dendrogram of hierarchical clustering of universities using ranks in OKI indicators, with universities coloured by university network.

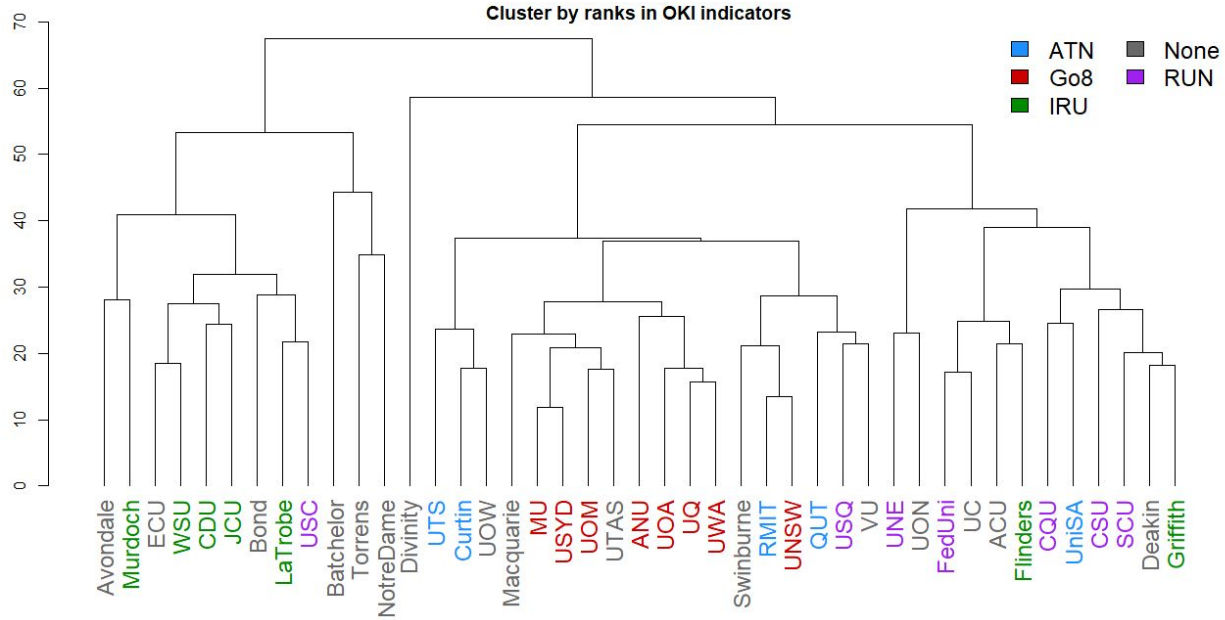
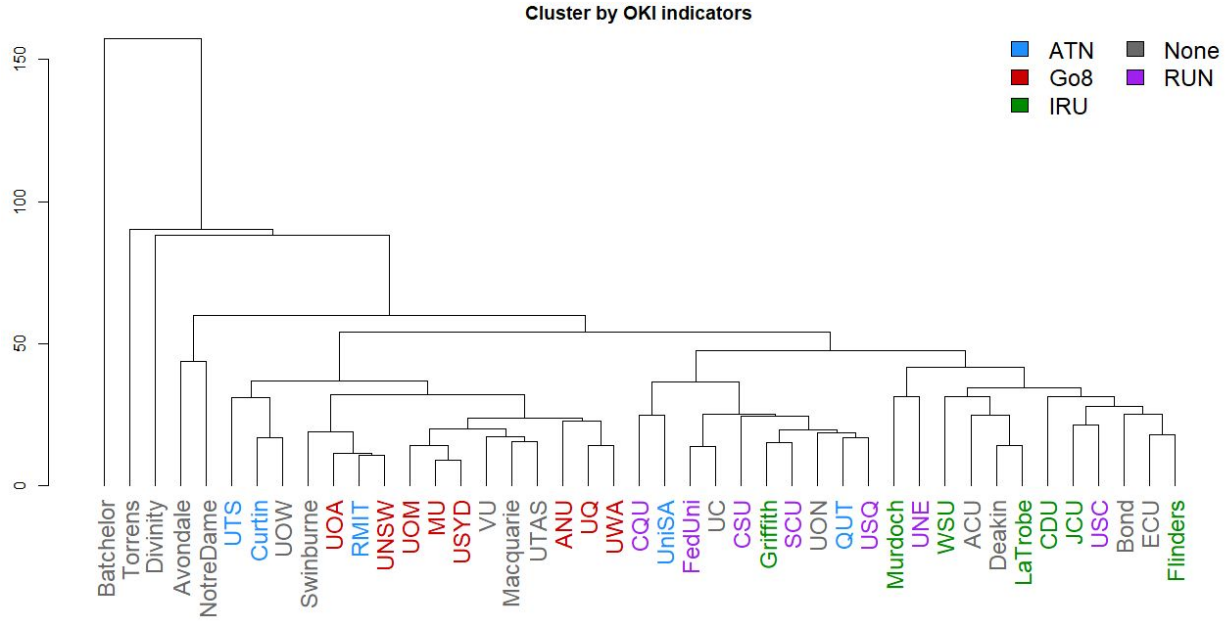


Figure 17: Dendrogram of hierarchical clustering of universities using OKI indicators, with universities coloured by university network.



4. Discussion on main findings

In this study we examined patterns and potential relationships across a number of OKI indicators for 43 Australian universities. We also explored ways in which information provided by these indicators can be summarised into a smaller number of orthogonal variables (PCs), and how they are aligned with the evaluation framework proposed by Montgomery et al. (2020). Universities are also clustered by using their corresponding data, ranks, and the corresponding PCs, to reveal overall similarities across universities based on the OKI indicators. In addition, an indicator of total revenue is also included in parts of our analysis for comparison. The main findings are summaries in a number of points below:

- Many of the indicators are characterised by high levels of disparity regardless of whether extreme observations are removed or not. The most severe cases (relatively) are “indigenous”, “total_rev” and “walk_score” when we focus on the middle 50% of observations for each indicator. Indicator “oa_gold” and some of the gender indicators also stand out when all observations are included. These signal the large differences across universities in Australia in terms of the characteristics related to these indicators.
- Combined with high levels of skewness and differences in measurement scale, the above implies a need for robust statistical methods when analysing such data. These robust methods need to cater for extreme observations (or outliers) and be able to standardise information across different variables. Hence, we used a diverse set of descriptive statistics, Spearman’s rank correlation, robust PCA and robust clustering methods for our analysis.

- Using pairwise Spearman's rank correlation, we found that Australian universities have a relatively larger focus on the green route for OA than other forms of OA. This is evidenced by a much higher correlation between "oa_total" and "oa_green" relative to the pairing of "oa_total" against other OA indicators. OA indicators are also negatively correlated to "output_div" and a potential explanation is the lack of complete information on the OA status of output formats other than journal articles. We have also found that publisher-mediated OA is not correlated to the university's total revenue!
- There is evidence of a general location effect on the diversity indicators. There are high correlations across "walk_score", "indigenous" and the gender indicators, albeit in different directions. Universities in more regional areas (lower "walk_score" in general) tend to have higher scores for "indigenous" and the gender indicators. They also have higher scores for "collab_aus" and lower scores for other collaboration indicators. The diversity indicators also display high negative correlation with "total_rev" in general.
- Low levels of variance explained by PCs and the high number of PCs needed to explain a significant portion of the total variance implies that the information provided by these OKI indicators are high-dimensional and complex. This is in contrast to the findings related to popular university rankings where only 2 or 3 PCs are needed to capture most of the total variance.
- We also show that the first 3 PCs after rotation can each be identified with groupings of indicators in diversity, communication, coordination, respectively. This is consistent with our analysis using the network plot of Spearman's rank correlation matrix. These form a proof-of-concept of the OKI evaluation framework proposed by Montgomery et al. (2020).
- Our cluster analysis using both PCs and using robust hierarchical clustering demonstrate evidence of greater alignment (in respect to the OKI indicators) within university networks than within geolocations. Such synergy is most strong among members within the Go8.

5. Limitations

While this study aims to be as comprehensive as possible in terms of both data collection and data analysis, there exist a number of limitations that need to be noted. Research outputs counted as part of this work are limited to those with existing DOIs that maps with our Crossref data snapshot. Hence, a research output without a recorded Crossref DOI is left out. Affiliation between research outputs and universities comes from three potential sources: Microsoft Academic, Web of Science, and Scopus. Each of these sources have their own limitations in accurately recording affiliation information. Readers are referred to Huang et al. (2020a) for a report on such limitations. It should be noted that these bibliographic data sources are also dynamic (i.e., continuously changing), including potential backfilling.

There exist missing values in the data that we have collected for the set of indicators. We have implemented and trialled various methods for the robust handling of these missing values. The

results obtained are largely consistent across different methods. However, we cannot discard the possibility for the (unobserved) real data associated with these missing values to be vastly different and can potentially change the results significantly. Similarly, there are a number of extreme observations in the data, including those driven by sample sizes. While we have used robust methods to counter the potential size effects of these extremes, they nevertheless remain the largest or smallest values within the respective set of indicator observations.

Extensive reviewing and manual work was used for the document analysis of university policies and statements for constructing the policy indicators. However, we cannot completely discard the possible subjectivities on our part of determining scores for these indicators. The annual report analysis is tested on two separate machines using different Python versions, and have yielded similar general results. However, we do note that the process of transforming PDF documents to text files can depend on the operating system and versions of softwares used.

Lastly, our analysis is focused on the small sample of Australian universities for a specific year. It remains unknown whether the main findings can be generalised to a larger region or over a longitudinal data set. There also remain challenges for collecting data over larger scales.

6. Conclusion

In this study, we introduced a selected number of OKI indicators that were collected by the COKI project based at Curtin University. We also explored several techniques for analysing these OKI indicators that are robust towards missing values, extreme observations and different measurement scales. The findings suggest significant disparities across Australia in terms of certain characteristics, such as levels of indigenous employment and gender equity. There is also evidence of a strong focus on repository-mediated access when it comes to OA provision of research outputs. Overall, the OKI indicators provide high dimensional and complex signals, which can be largely grouped into three categories of diversity, communication and coordination. This is largely in agreement with the OKI evaluation framework described by Montgomery et al. (2020).

Data Sharing

The codes and the relevant curated data for analysis are made available through Zenodo at https://doi.org/10.5281/zenodo.4040402. However, the raw data related to publications are not shared to respect the terms of service of the data sources. Links to codes and data for the policy analysis and the annual report analysis are also provided in the respective appendices.

Acknowledgements

This work is funded by the Research Office of Curtin University through a strategic grant, the Curtin University Faculty of Humanities, and the School of Media, Creative Arts and Social Inquiry.

References

Altunkaynak, B., Gamgam, H. (2019). Bootstrap confidence intervals for the coefficient of quartile variation. *Communications in Statistics - Simulation and Computation* 48(7): 2138-2146. https://doi.org/10.1080/03610918.2018.1435800

Australian Academic of Humanities (2015). Measuring the Value of International Research Collaboration. https://www.humanities.org.au/wp-content/uploads/2017/04/AAH_Measuring-Value-2015.pdf

Baker, M. (2016). Women graduates and the workplace: continuing challenges for academic women. *Studies in Higher Education* 41(5): 887-900. https://doi.org/10.1080/03075079.2016.1147718

Bonett, D. G. (2006). Confidence interval for a coefficient of quartile variation. *Computational Statistics & Data Analysis* 50(11): 2953-2957. https://doi.org/10.1016/j.csda.2005.05.007

Dehon, C., McCathie, A., Verardi, V. (2010). Uncovering excellence in academic rankings: a closer look at the Shanghai ranking. *Scientometrics* 83: 515–524. https://doi.org/10.1007/s11192-009-0076-0

Docampo, D. (2011). On using the Shanghai ranking to assess the research performance of university systems. *Scientometrics* 86: 77–92. https://doi.org/10.1007/s11192-010-0280-y

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. Scientometrics 51: 69-115. https://doi.org/10.1023/A:1010512628145

Glänzel, W., de Lange, C. (2002). A distributional approach to multinationality measures of international scientific collaboration. Scientometrics 54: 75-87. https://doi.org/10.1023/A:1015684505035

Goglio, V. (2016). One size fits all? A different perspective on university rankings. *Journal of Higher Education Policy and Management* 38(2): 212-226. https://doi.org/10.1080/1360080X.2016.1150553

Hazelkorn, E. (2008). Learning to live with league tables and ranking: the experience of institutional leaders. *Higher Education Policy* 21: 193–215. https://doi.org/10.1057/hep.2008.1

Hazelkorn, E. (2009). Rankings and the Battle for World-Class Excellence: Institutional Strategies and Policy Choices. *Higher Education Management and Policy* 21(1): 55-77. https://www.oecd-ilibrary.org/content/paper/hemp-v21-art4-en

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., Ozaygen, A. (2020a). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies* 1(2): 445-478. https://doi.org/10.1162/qss_a_00031

Huang, C-K., Neylon, C., Hosking, R., Montgomery, L., Wilson, K., Ozaygen, A., Brookes-Kenworthy, C. (2020b). Evaluating institutional open access performance: Methodology, challenges and assessment. *bioRxiv*. https://doi.org/10.1101/2020.03.19.998336

Hubert, M., Rousseeuw, P., Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics* 47(1): 64-79. https://doi.org/10.1198/004017004000000563

Johnes, J. (2018). University rankings: What do they really show? *Scientometrics* 115, 585-606. https://doi.org/10.1007/s11192-018-2666-1

Kader, G. D., Perry, M. (2007). Variability for Categorical Variables. *Journal of Statistics Education* 15(2). https://doi.org/10.1080/10691898.2007.11889465

Kehm, B. M. (2014). Global university rankings – impacts and unintended side effects. *European Journal of Education* 49(1): 102–112. https://doi.org/10.1111/ejed.12064

Khan, M. S., Lakha, F., Tan, M. M. J., Singh, S. R., Quek, R. Y. C., Han, E., Tan, S. M., Haldane, V., Gea-Sánchez, M., Legido-Quigley, H. (2019). More talk than action: gender and ethnic diversity in leading public health universities. *The Lancet* 393(10171): 594-600. https://doi.org/10.1016/S0140-6736(18)32609-6

Marini, G., Meschitti, V. (2018). The trench warfare of gender discrimination: evidence from academic promotions to full professor in Italy. *Scientometrics* 115: 989-1006. https://doi.org/10.1007/s11192-018-2696-8

Montgomery, L., Hartley, J., Neylon, C., Gillies, M., Gray, E., Herrmann-Pillath, C., Huang, C-K., Leach, J., Potts, J., Ren, X., Skinner, K., Sugimoto, C., Wilson, K. (2020 In press), *Open Knowledge Institutions: Reinventing Universities*. Cambridge MA, MIT Press.

Niles, M. T., Schimanski, L. A., McKiernan, E. C., Alperin, J. P.. (2020). Why we publish where we do: Faculty publishing values and their relationship to review, promotion and tenure expectations. *PLoS ONE* 15(3): e0228914. https://doi.org/10.1371/journal.pone.0228914

Olejniczak, A., Wilson, M. J. (2020). Who's writing Open Access (OA) articles? Characteristics of OA authors at Ph.D. granting institutions in the USA. *SocArXiv*. https://doi.org/10.31235/osf.io/gcr32

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 6:e4375. https://doi.org/10.7717/peerj.4375

Selten, F., Neylon, C., Huang, C-K., Groth, P. (2020). A longitudinal analysis of university rankings. *Quantitative Science Studies* 1(3): 1109-1135. https://doi.org/10.1162/qss_a_00052

Siler, K., Haustein, S., Smith, E., Larivière, V., Alperin, J. P. (2018). Authorial and institutional stratification in open access publishing: the case of global health research. *PeerJ* 6:e4269. https://doi.org/10.7717/peerj.4269

Subbaye, R., Vithal, R. (2017). Gender, teaching and academic promotions in higher education. *Gender and Education* 29(7): 926-951. https://doi.org/10.1080/09540253.2016.1184237

Wilson, K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Huang, C-K., Montgomery, L., Ozaygen, A. (2019a). 'Is the library open?': Correlating unaffiliated access to academic libraries with open access support. *LIBER Quarterly* 29(1): 1-33. http://doi.org/10.18352/lq.10298

Wilson, K., Neylon, C., Montgomery, L., Huang, C-K. (2019b). Access to academic libraries: An indicator of openness? *Information Research* 24(1): paper 809. http://InformationR.net/ir/24-1/paper809.html

Wilson, K., Neylon, C., Montgomery, L., Huang, C-K., Hosking, R., Ozaygen, A. (2020). Global diversity in higher education staffing: Towards openness. http://dx.doi.org/10.17613/tyhv-h252

Winslow, S., Davis, S. N. (2016). Gender Inequality Across the Academic Life Course. *Sociology Compass* 10(5): 404-416. https://doi.org/10.1111/soc4.12372

Appendix A: Justification for the “event_total” definition

To be consistent with the construction of other indicators relating to research outputs, we have selected to base the “event_total” indicator on counting the number of publications satisfying the condition of having at least one Crossref event. However, the choice of “at least one” needs some justification. Hence, we take a closer look at how the universities’ performances change when we adjust the condition of “at least one” to “at least 2”, “at least 3”, etc. The results are summarised in Figure A1.

Figure A1: Proportion of outputs with specified minimum numbers of Crossref events, for all 43 universities.

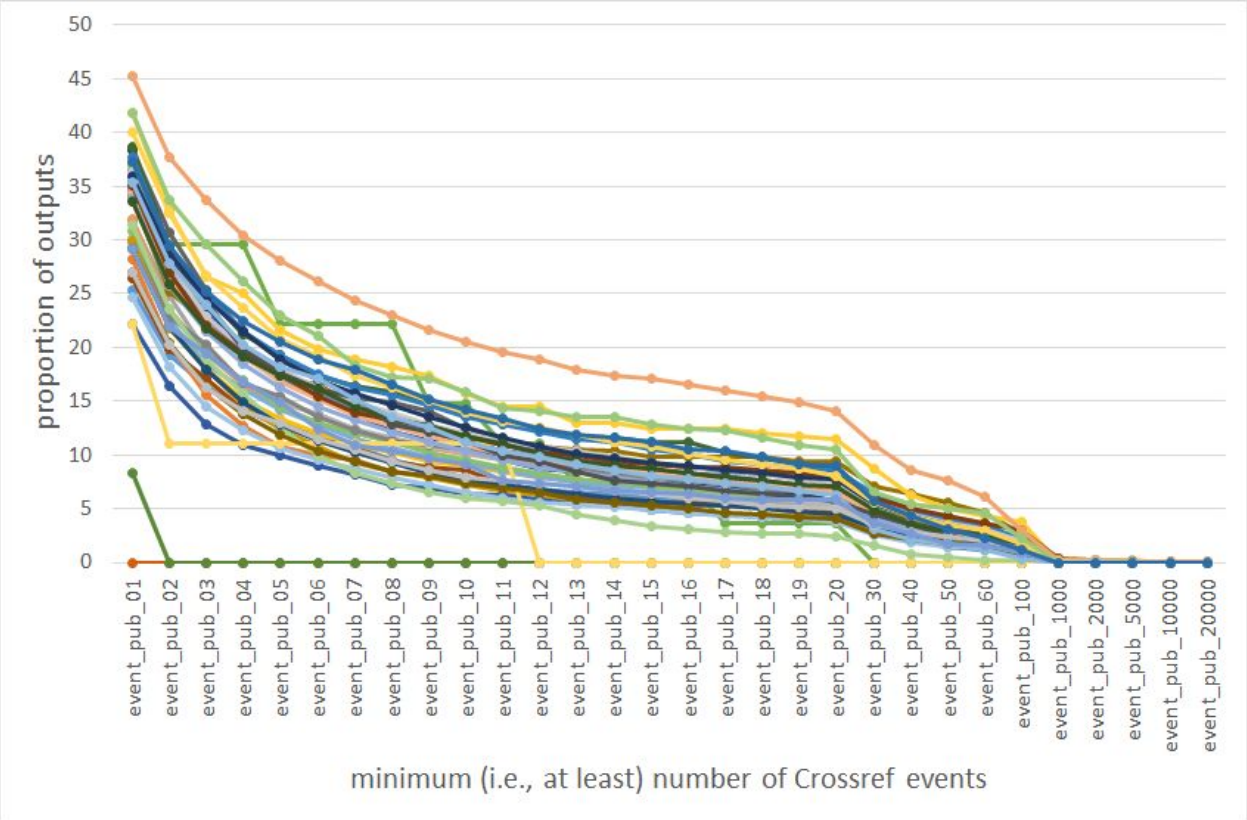


Figure A1 records, for each university (represented by each line), the proportion of its outputs meeting the conditions of having at least some number of Crossref events. We observe exponential decreases in the proportions as the requirement for minimum number of Crossref events increases. This is not unexpected given the exponential increase in number of events when outputs are arranged in order of number of events, i.e., there is a large number of outputs

with 0 events, followed by one event, and so on, with very few number of outputs associated with extremely high numbers of events.

We note that there exist a few large jumps. These are representative of universities with very small numbers of outputs. Apart from these, the decreases in proportions of outputs generally appears to be in parallel. Hence, the relative performance (i.e., ranks) across the universities does not appear to drastically change due to changes to the required minimum number of events. As a result, we use “at least one” for simplicity.

We have also avoided using the actual value of event counts at the output level for constructing this indicator, such as the average number of events per article. Such a measure is highly influenced by outliers. For example, a small university in our data has one publication with more than 10000 tweets, significantly higher than the rest of its outputs. This publication includes more than 15000 signatories from other scientists, which may have had an impact on the number of tweets. This single output resulted in the university to have a much higher average event count than all other universities, albeit only having a fraction of number of output compared to other highly ranked universities.

This is an example of the general outlier problem for such data. If such multiplicative nature is persistent in the data, then an alternative measure of central tendency may be appropriate, e.g., geometric mean. But the extent to which such an approach can alleviate the effects of these observations is yet unknown. Hence, we have decided not to aggregate the number of events due to this issue.

Appendix B: Data collection on policies and infrastructure

We gathered documents from university public websites supplemented by directories and collections such as the Directory of Open Access Repositories (DOAR) (http://v2.sherpa.ac.uk/opendoar/), the Registry of Open Access Repository Mandates and Policies (ROARMAP) (https://roarmap.eprints.org/) and Politicas MELIBEA (https://www.accesoabierto.net/politicas/), a directory and estimator of OA policies for institutional repositories and practices. We developed a user-assisted tool to automate the search, retrieval and downloading of library access policy, OA policy and diversity policy documents from university websites. The tool consists of a Jupyter notebook supported by a small library of Python code. Using the Bing search engine API it executes a search against the URL for a specific university website recorded in the Global Research Identifier Database - GRID (https://www.grid.ac/). The code and an example Jupyter notebook are available at https://doi.org/10.5281/zenodo.1438874. This process was supplemented with human search and retrieval where necessary.

Subsequently, information on policies surrounding library access, OA and diversity were manually retrieved and used to answer a number of Yes/No questions (1=Yes; 0=No). These are then used to construct respective indicators in the following way:

- Library public access score ("policy_lib", score out of 3):
    - Is the library accessible by the public?
    - Is the library freely (i.e., no fee) accessible by the public?
    - Is the library accessible without restrictions (e.g., ID requirements)?
- OA score ("policy_div", score out of 5):
    - Does the university have an OA policy or statement?
    - Does the university provide extra funding for OA publishing?
    - Does the university have an OA repository?
    - Does the university have a data sharing policy or statement?
    - Does the university have an open data repository?
- Diversity policy score ("policy_div", score out of 2):
    - Does the university have a policy on employment equity, equality or equity?
    - Does the university have a policy on staff diversity?

The process of retrieving policy documents and related information involved a vast amount of manual work, including several reviews of documents, manual searches and examining multiple weblinks. This process is described in detail in Wilson et al. (2019a, 2019b, 2020). The policy documents examined are copyright as per the respective universities and are used here for study and research only.

Appendix C: Data collection and analysis of annual reports

The individual annual reports of the universities are manually downloaded (if publicly available). A Python script is used to convert these PDF files into text file format, convert words in the documents into tokens, and identify a set of predefined phrases in each text document and record the number of times each phrase appears in the documents. The phrases are grouped into the three platforms of diversity, communication and coordination, with the aggregated relative frequency (out of total number of words in the document) used for the respective OKI indicators.

The downloaded annual reports and python scripts used for text analysis are made available on Zenodo (https://doi.org/10.5281/zenodo.4034821). The words used and the groupings can be found in the JSON file "words.json", with the respective final word counts and relative frequencies recorded in "AU_2017.csv".

The annual report PDF files are copyright as per the respective universities, where applicable. They are used here for study and research purposes only.

Appendix D: Additional outputs for descriptive analysis

Table D1 provides a number of summary statistics for each of the indicators, along with the number of missing values within each indicator.

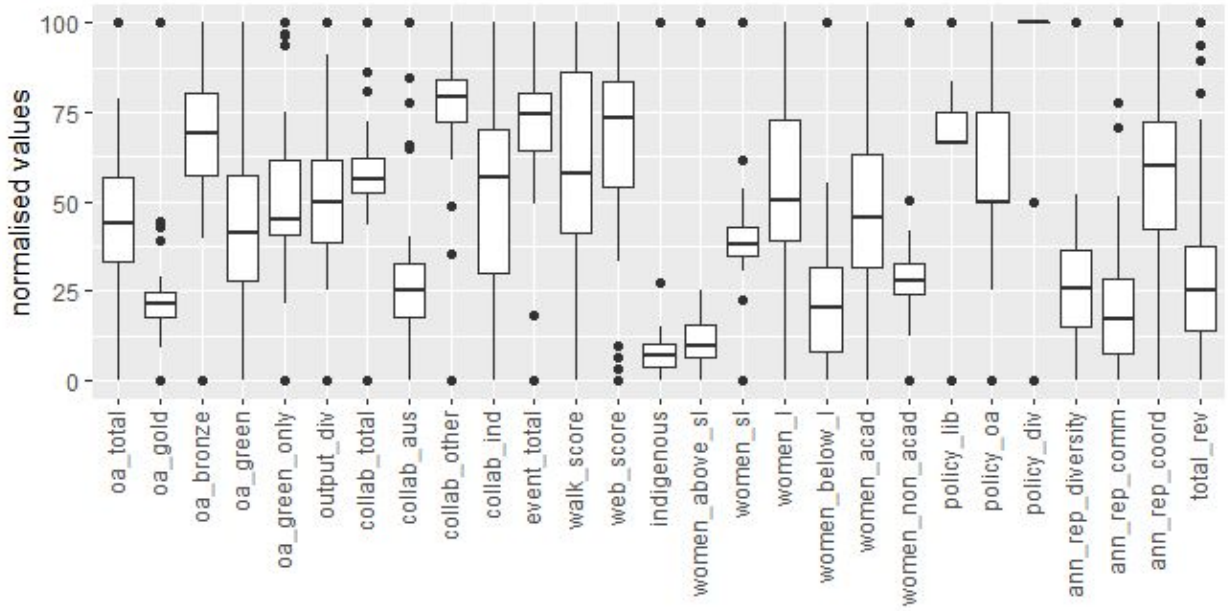
Table D1: Summary statistics of indicators.

	Min	1st Qu	Mean	Median	3rd Qu	Max	NA
oa_total	25.00	38.82	43.88	43.37	48.60	66.67	0
oa_gold	8.33	18.65	22.01	20.77	22.67	66.67	0
oa_bronze	0.00	6.93	7.89	8.33	9.66	12.04	0
oa_green	25.00	32.73	36.79	36.44	40.90	52.75	0
oa_green_only	0.00	11.27	13.98	12.40	17.00	27.64	0
output_div	0.00	0.17	0.23	0.22	0.27	0.44	0
collab_total	41.67	72.39	75.23	74.61	77.79	100.00	0
collab_aus	26.74	39.67	48.19	45.13	50.62	100.00	0
collab_other	0.00	51.15	53.24	55.87	59.34	70.67	0
collab_ind	2.30	3.20	3.87	4.00	4.40	5.30	17
event_total	0.00	29.18	31.65	33.62	36.38	45.27	0
walk_score	7.00	45.50	62.77	61.00	87.00	100.00	0
web_score	21.00	37.25	40.71	43.00	46.00	51.00	1
indigenous	0.00	0.67	1.60	1.12	1.66	16.13	0
women_above_sl	26.03	30.83	35.30	33.10	37.69	100.00	1
women_sl	13.33	43.75	47.54	46.30	50.56	100.00	2
women_l	42.11	52.11	55.70	54.89	60.65	67.57	2
women_below_l	44.13	48.71	56.67	55.56	61.75	100.00	3
women_acad	35.68	43.64	47.26	47.04	51.40	60.61	4
women_non_acad	53.85	65.00	67.40	66.77	68.87	100.00	1
policy_lib	0.00	2.00	2.15	2.00	2.25	3.00	0
policy_oa	1.00	3.00	3.23	3.00	4.00	5.00	0

policy_div	0.00	2.00	1.74	2.00	2.00	2.00	0
ann_rep_diversity	0.0017	0.0027	0.0034	0.0033	0.0040	0.0080	9
ann_rep_comm	0.0002	0.0004	0.0007	0.0006	0.0008	0.0024	9
ann_rep_coord	0.0028	0.0045	0.0052	0.0053	0.0058	0.0069	9
total_rev	40742.00	380714.50	821233.10	664774.00	961264.50	2501975.00	4

To highlight the extreme points for each OKI indicator and to allow for cross-comparison of these indicators, we normalise[19] each indicator using the min-max rescaling. This results in each indicator being rescaled to the range between 0 and 100. The boxplots of these normalised indicators are presented in Figure D1. The dots represent extreme observations[20].

Figure D1: Boxplots of normalised observations for OKI indicators.



[19] Each indicator is normalised using the min-max rescaling, i.e., normalised value = (original value - min) / (max - min), where min and max are the minimum and maximum of all values observed for the given indicator.

[20] Observations that lie more than 1.5 times the interquartile range away from the first or third quartile.

Appendix E: Additional PCA and cluster analysis results

Figure E1: Scree plot of eigenvalues with Kaiser criterion for PCA with Spearman's rank correlation (left) and ROBPCA (right).

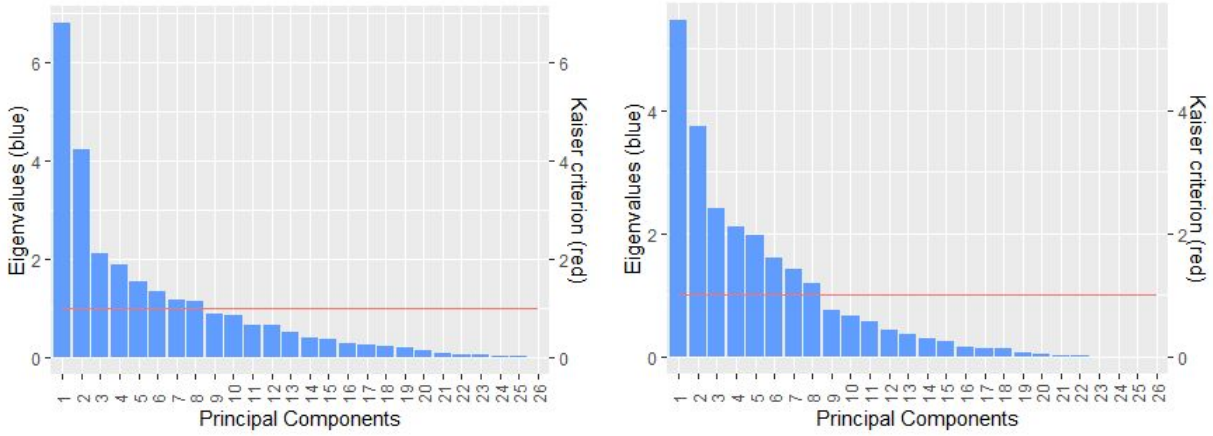


Table E1: Standardised loadings on the first 8 PCs from the Spearman PCA.

Indicators	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
oa_total	-0.06	0.38	-0.01	0.24	0.22	-0.01	0.14	-0.02
oa_gold	0.12	0.37	-0.05	-0.02	-0.18	0.08	-0.16	0.00
oa_bronze	-0.21	0.14	0.09	-0.32	0.00	0.28	0.12	0.03
oa_green	-0.11	0.39	-0.03	0.24	0.18	-0.10	0.16	0.08
oa_green_only	-0.17	0.21	-0.01	0.24	0.26	-0.22	0.30	0.16
output_div	0.01	-0.36	0.04	0.17	0.11	-0.18	0.12	-0.02
collab_total	0.06	0.33	-0.29	-0.09	0.02	-0.11	-0.21	-0.18
collab_aus	0.31	0.10	-0.07	-0.13	-0.05	-0.04	0.01	0.07
collab_other	-0.23	0.28	-0.11	-0.11	0.04	0.06	-0.16	-0.22
collab_ind	-0.20	0.11	0.04	0.08	-0.32	0.14	0.42	-0.22
event_total	0.09	0.27	0.35	-0.10	-0.17	-0.10	0.06	0.26
walk_score	-0.21	0.04	-0.33	-0.34	0.10	-0.01	-0.07	-0.12

web_score	-0.02	-0.01	0.24	-0.34	0.20	-0.03	0.22	0.45
indigenous	0.17	0.02	0.20	0.45	0.02	0.07	-0.08	-0.28
women_above_sl	0.32	0.00	-0.04	-0.16	0.12	-0.14	-0.06	-0.03
women_sl	0.31	0.13	-0.02	-0.09	0.05	-0.19	-0.01	-0.08
women_l	0.33	0.13	0.08	0.03	0.10	-0.06	0.00	0.05
women_below_l	0.29	0.05	-0.03	0.04	-0.18	0.11	0.23	0.04
women_acad	0.36	0.09	0.03	-0.04	-0.01	0.00	0.08	-0.03
women_non_acad	0.23	0.01	0.00	-0.12	0.11	0.09	0.07	-0.28
policy_lib	0.10	-0.03	0.11	0.16	0.20	0.66	0.01	-0.07
policy_oa	-0.10	0.19	0.46	-0.08	-0.18	0.09	-0.24	0.02
policy_div	-0.08	0.01	0.46	-0.13	0.26	0.00	-0.30	-0.22
ann_rep_diversity	-0.11	-0.04	0.29	-0.04	0.16	-0.45	0.03	-0.36
ann_rep_comm	0.05	-0.01	-0.11	0.02	0.60	0.21	-0.13	0.18
ann_rep_coord	0.08	-0.01	0.06	-0.33	0.17	0.08	0.52	-0.40

Table E2: Standardised loadings on the first 7 PCs from ROBPCA.

Indicators	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
oa_total	-0.15	-0.41	0.03	0.04	0.03	0.13	0.13
oa_gold	-0.21	-0.13	-0.04	-0.01	-0.05	-0.02	-0.04
oa_bronze	0.00	-0.18	0.00	0.20	-0.35	-0.21	-0.13
oa_green	-0.13	-0.48	0.03	-0.01	0.14	0.21	0.18
oa_green_only	0.10	-0.25	0.10	-0.03	0.27	0.29	0.28
output_div	0.34	0.27	0.00	-0.11	0.09	0.24	0.05
collab_total	-0.27	-0.09	0.18	0.00	-0.06	-0.09	-0.12
collab_aus	-0.39	0.14	0.14	-0.03	0.03	0.04	-0.09
collab_other	-0.04	-0.24	0.17	0.07	-0.09	-0.16	-0.10
collab_ind	0.13	-0.23	0.09	0.03	-0.31	0.11	0.34

event_total	-0.32	-0.08	-0.21	-0.07	-0.11	0.03	-0.01
walk_score	0.07	-0.12	0.28	0.40	-0.27	-0.08	-0.24
web_score	0.11	0.09	-0.53	0.31	-0.32	-0.10	0.46
indigenous	-0.05	0.02	-0.02	-0.10	0.10	0.04	-0.04
women_above_sl	-0.20	0.14	-0.02	0.02	-0.05	0.01	-0.04
women_sl	-0.27	0.08	-0.07	-0.05	-0.12	0.10	0.02
women_l	-0.36	0.09	-0.17	-0.02	0.03	0.08	0.04
women_below_l	-0.21	0.17	-0.04	0.05	0.00	0.15	0.21
women_acad	-0.29	0.17	-0.10	-0.01	-0.06	0.08	0.05
women_non_acad	-0.08	0.08	0.03	0.06	-0.02	0.05	-0.06
policy_lib	-0.03	0.12	0.03	0.19	0.18	-0.17	-0.02
policy_oa	-0.02	-0.28	-0.40	-0.31	-0.11	-0.11	-0.13
policy_div	0.17	-0.12	-0.41	-0.12	-0.02	-0.04	-0.42
ann_rep_diversity	0.09	-0.11	-0.16	-0.18	0.04	0.34	-0.30
ann_rep_comm	-0.06	-0.12	-0.30	0.60	0.56	-0.05	-0.12
ann_rep_coord	0.02	0.07	-0.03	0.33	-0.28	0.69	-0.27

Table E3: Rotated loadings on first 3 PCs from Spearman PCA and ROBPCA.

	Spearman PCA			**ROBPCA**		
Indicators	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
oa_total	0.0109	**0.3864**	0.0238	0.0065	**-0.4351**	-0.0563
oa_gold	0.1902	**0.3421**	-0.0243	-0.1602	-0.1953	-0.0386
oa_bronze	-0.1717	0.1703	0.1055	0.0602	-0.1625	-0.0445
oa_green	-0.0339	**0.4057**	0.0049	0.0521	**-0.4936**	-0.0795
oa_green_only	-0.1297	0.2425	0.0154	0.2053	**-0.2073**	0.0027
output_div	-0.0635	**-0.3589**	0.0089	0.2167	0.376	0.0145
collab_total	0.1128	**0.3403**	-0.2672	-0.1656	**-0.2224**	0.191
collab_aus	**0.3225**	0.0486	-0.0725	**-0.3655**	-0.0514	0.2436

collab_other	-0.1727	**0.3294**	-0.081	0.0898	**-0.2659**	0.0974
collab_ind	-0.1689	0.1459	0.0541	0.2206	-0.1703	0.0006
event_total	0.1528	0.2146	**0.3684**	**-0.3148**	-0.1583	-0.1703
walk_score	-0.2091	0.1048	**-0.3175**	0.1823	-0.1358	0.2174
web_score	-0.0136	-0.0255	0.2427	-0.0768	0.2265	**-0.4968**
indigenous	0.178	-0.0342	0.1933	-0.0574	0.0007	-0.0086
women_above_sl	**0.3146**	-0.0625	-0.0573	-0.2317	0.0551	0.0511
women_sl	**0.3234**	0.0659	-0.0216	**-0.2917**	-0.0169	0.0082
women_l	**0.3489**	0.0529	0.0754	**-0.4042**	-0.0272	-0.071
women_below_l	**0.2949**	-0.0027	-0.0354	**-0.2519**	0.0803	0.0481
women_acad	**0.3737**	0.0147	0.0195	**-0.3454**	0.0564	0.0012
women_non_acad	0.2253	-0.0375	-0.0142	-0.0916	0.0305	0.0676
policy_lib	0.0995	-0.0576	0.1069	-0.0628	0.0903	0.0687
policy_oa	-0.0523	0.1632	**0.4806**	-0.028	-0.186	**-0.4547**
policy_div	-0.0614	-0.0173	**0.4627**	0.0867	0.0374	**-0.449**
ann_rep_diversity	-0.1093	-0.0396	**0.2901**	0.0765	-0.0358	-0.1987
ann_rep_comm	0.0399	-0.0081	-0.1146	-0.0895	-0.0729	**-0.3018**
ann_rep_coord	0.0746	-0.0321	0.058	-0.0116	0.0784	-0.0128

Figure E2: Individual component plot for first two PCs from ROBPCA, with universities coloured by state.

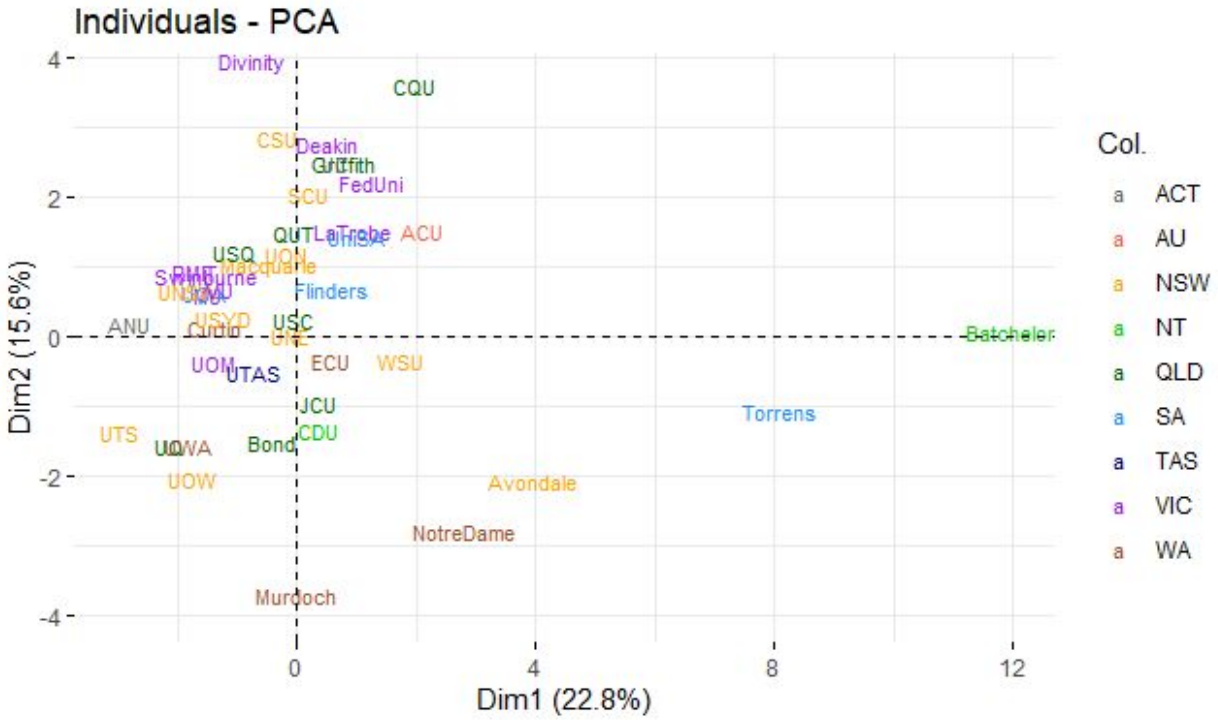


Figure E3: Individual component plot for first two PCs from ROBPCA, with universities coloured by university network.

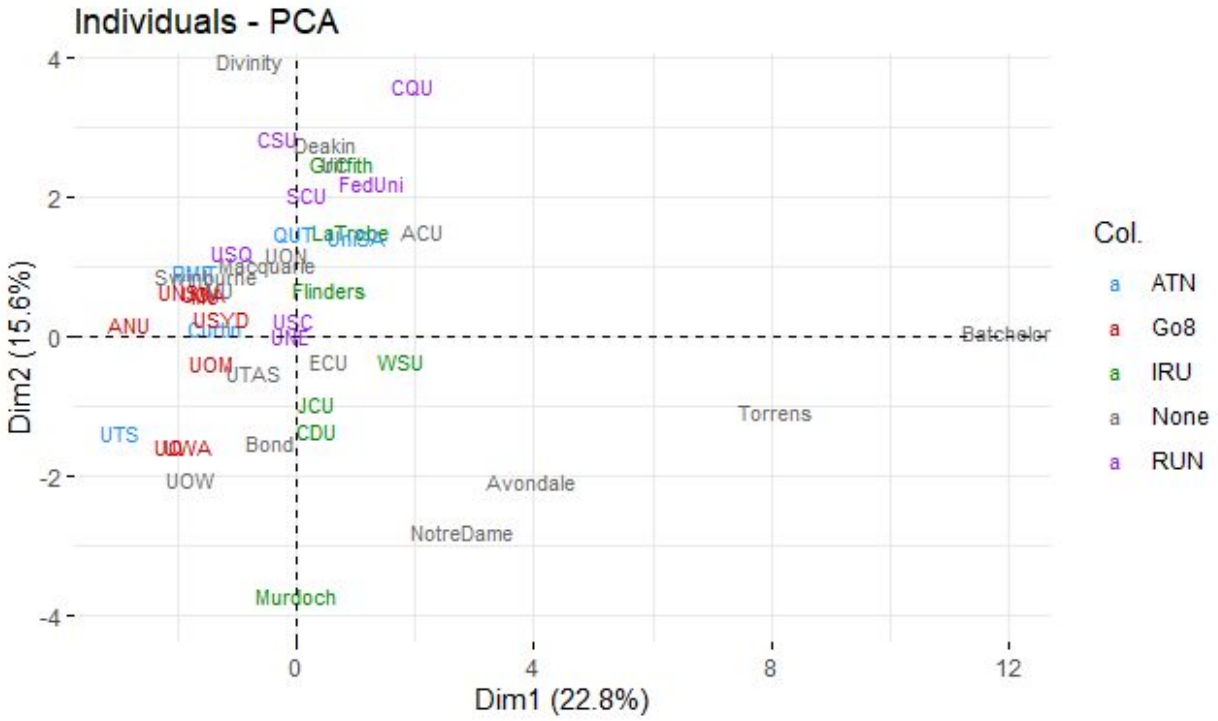


Figure E4: Dendrogram of hierarchical clustering of universities using ranks in OKI indicators, with universities coloured by state.

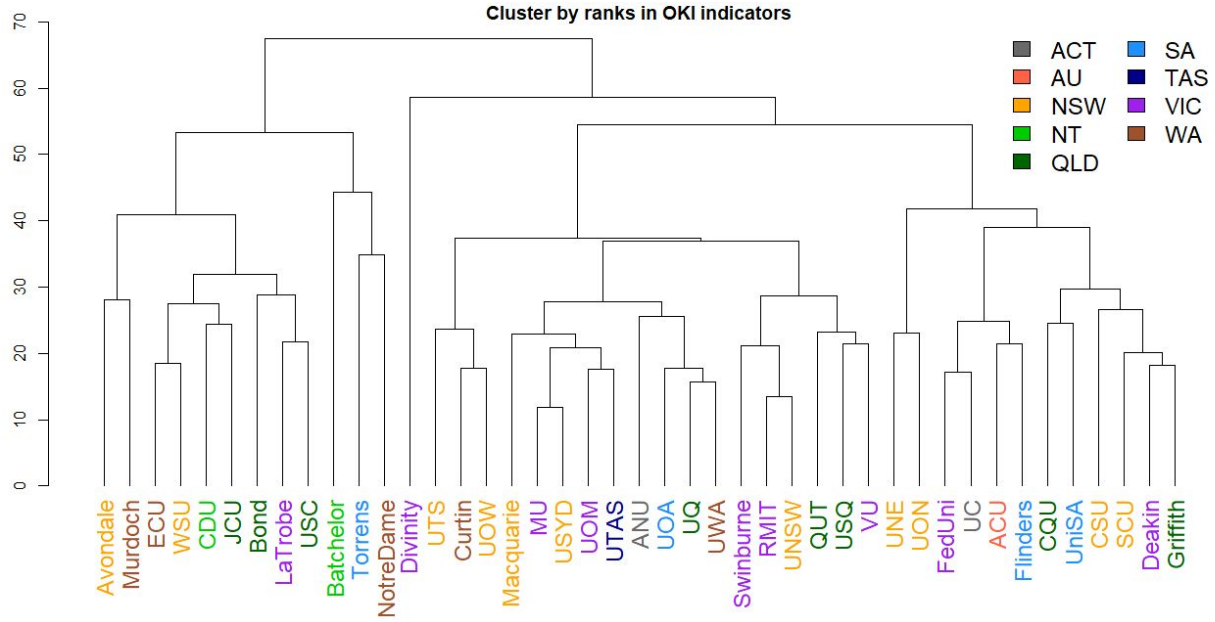


Figure E5: Dendrogram of hierarchical clustering of universities using OKI indicators, with universities coloured by state.