Co-authorship Networks could drive Citations

Adam Butler
Advanced Analytics Institute
University of Technology Sydney
Research Portfolio
University of Sydney
Adam.Butler@sydney.edu.au

Guandong Xu
Advanced Analytics Institute
University of Technology Sydney
Guandong.Xu@uts.edu.au

ABSTRACT

For many decades citation counting has been used as the way to quantify the nebulous notion of research "quality". Indeed, in conversation the terms "research quality", "impact" or "excellence in research" are simply a reference to a scientific document's citation count. Moreover, the commonly used journal "impact" factors are simply manipulated forms of citation counting. In recent times, the word "impact" has morphed into the new '*mot du jour*'.

This paper investigates and discusses the association between co-authorship networks and citations of institutions within an arbitrary, but defined, subject area. The data examined is readily available and the analytical techniques employed are deliberately simple. The simplicity of this analysis is driven by the desire to show that citation counts are not explicitly related to the quality of research but that citations are a result of multifaceted author networks that are inherent in scientific endeavor.

The paper presents an argument that the improved ability to conduct effective network analysis and related research shows that the notion of high citations being the same as "research quality" has run its course. Citation performance is more likely to be a result of co-authorship network dynamics rather than any perceived notion of "quality". Moreover, it is time the folly of citation counting is put to rest and that if one wants know what "impact" one is having that you need look no further than your co-authorship network and the reach it has across whatever subject area you are interested in.

The discussion and results herein highlight that rather than counting citations, the "impact" of research is driven by connections through networks of people.

Keywords

Co-authorship networks, citation analysis, research quality, research impact

Introduction

It seems the obsession with citation counting is ubiquitous throughout academia. Researchers are more often than not judged on their citation average and/or the plethora of other citation-based metrics such as h-index and variants thereof [1-3]. The terms "research quality", "research impact" and "research excellence" have become synonymous with citation counting. But, is citation counting really an indication of an individual researcher's worth or the "quality" of their work? Or indeed, is the citation average of an institution really the sum part of that institution's worth and quality? The reader is likely aware of the multitude of international university ranking systems that use citation counting to inform their published rankings. This is to say nothing of national government-led research assessment schemes such as one in Australia called the 'Excellence in Research for Australia' (ERA) exercise which draws upon citation counting as a measure of quality.

The situation of using citations as a proxy for quality can perhaps trace its origins to a paper by Garfield [4] that described compiling citation lists in order to improve the efficiency of finding relevant research. In that paper Garfield writes that "in the case of a highly significant article, the citation index has a quantitative value,

for it may help the historian to measure the influence of the article – that is its impact factor."

Citation counting as a way to assess researchers has long been controversial to the extent that Garfield [5] rejects critics of citation analysis to rate scientific performance. But Garfield is correct in his observation that as the "scientific enterprise becomes larger and more complex, and its role in society more critical, it will become more difficult, expensive and necessary to evaluate and identify the largest contributors." Garfield suggests that citation analysis' virtue lies in its "relatively low cost" [5]. It would appear from Wade [6] that acceptance of citation analysis would only be a matter of time due to the growing consensus to its use as a performance measure.

However, as readers would appreciate, the scientific endeavor means that old ideas are tested and potentially lead to progress and the evolution of ideas. In this sense this is where co-authorship analysis is introduced. Of course, co-authorship analysis is not new and has been used to examine scientific collaborations for over a decade or more [7]. In some instances collaborations have been investigated in conjunction with citation data [8, 9]. Both Abbasi [8] and Biscaro [9] show positive correlations between scientific (co-authorship) network structure, position of authors and their citation performance. Given the evolution and current understanding of social network analysis and its derivatives, the question needs to be asked if it is time to throw citation analysis (as a proxy for quality) out the window? Given that some research has shown positive correlations between network structure and citations and that the networks (co-authorships) in question are mostly formed *before* citations are given, surely the networks drive the citations and not some vague notions of perceived "quality" and/or "excellence". As Garfield has acknowledged, scientific enterprise has become larger and much more complex with international collaborations increasing year on year. This paper contributes to the body of knowledge on co-authorship network analysis by considering the correlation between citation counts and simple network data. The paper refers to previous co-authorship studies that have shown relationships between networks and performance and proposes that networks dictate citations. Therefore it could be that citation counting has reached its use-by date.

In addition, all this talk of citations and quality says nothing of the 'Matthew Effect' in Science [10-13] where a "few countries with high expectations receive more citations than expected while many countries with low expectations receive fewer citations than expected."[11]

Method and Data

This paper explores simple network characteristics of a defined subject area taken from the widely used citation database Scopus. The subject area of interest, 'dentistry', was chosen because it is relatively narrow in scope. Moreover, 'dentistry' is a subject area defined in the QS university ranking system which also derives citation data from Scopus. This overlap is an important consideration in the approach taken for this paper since the chosen publication time period, 2011 to 2015, coincides with the time period for the 2016 QS ranking results.

The steps taken for data gathering, cleaning and analysis is outlined in Figure 1. Data was extracted from Scopus using the SciVal interface. Average citation and average relative citation based on year of publication (RCI_{year}) was calculated for each institution in the dataset.

The citation-based metrics were compared against the ratio of edges : degree in the 'ego' network of a select group of 50 institutions. The top 50 ranked institutions for dentistry in the 2016 QS system were chosen. Pearson and Spearman correlations were employed to test the null hypothesis (H_0) i.e. the ratio of edges : degree in the 'ego' network does not correlate with dataset institution-level citation averages. Spearman correlation was used because it tests monotonic relationships between two continuous or ordinal variables (based on the ranked values for each variable rather than the raw data).

One of the limitations often mentioned in co-authorship analysis is the identification of duplicate and/or erroneous affiliation data [14]. In Scopus, institutions are often associated with multiple affiliation IDs and this can make precise network node creation problematic. This issue, in turn, makes comparisons of data from different

origins somewhat challenging. For instance, there is no real way of knowing how Scopus cleans and/or consolidates multiple IDs in their data. Moreover, it is unknown how this data is aggregated and presented to ranking systems such as QS. This is a major hindrance. An example of an ambiguous affiliation is something like "School of Dentistry" which will, and does, have many affiliation IDs. When this occurs, consolidating and de-duplicating can produce great doubt in the mind of the analyst. De-duplication in the modified dataset was carried out manually when duplicate IDs were obviously associated with the same institution. If there was any doubt, the affiliation IDs were not altered.

In keeping with the desire for a simple and uncomplicated approach to the network data analysis, open-source network software Gephi was used to analyze network data. For this analysis, edges were weighted based on the number of institution affiliations associated with each document such that:

$$\text{edge weight} = \frac{1}{(n-1)}$$ where 'n' is the number of affiliations.

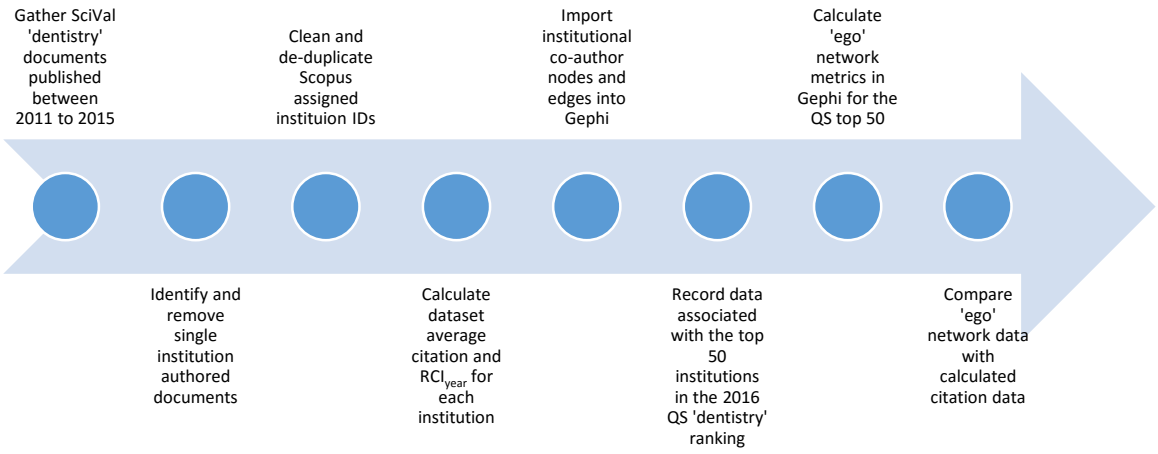

Figure 1 - Method steps

Approximately 6,112 institutions (72% of the nodes) were associated with just 1 or 2 publications in the modified dataset. The publication count 95th percentile was 24.

Data for the top 50 ranked institutions from the 2016 QS ranking in 'dentistry' was incorporated into the modified dataset. The QS "ranking indicators" (i.e. overall score, academic reputation, employer reputation and citations per paper) were compared against the resultant Gephi network attributes (centralities, between-ness etc…) for the 2016 QS top 50. For each of top 50 QS institutions their 'ego' networks were analyzed. The ratio of total edge count in the 'ego' network against the total direct connections (degree) was calculated.

Results and Discussion

There were 37,794 publications listed in SciVal within the subject area of 'dentistry' published during 2011 to 2015. Since we are interested in co-authorships between institutions, sole author publications and publications with only one associated institution were ignored and resulted in a modified dataset of 19,179 multi-institutional publications. These 19,179 publications gave rise to approximately 8,500 unique nodes (institutions), notwithstanding inaccuracies in de-duplication.

The de-duplicated nodes were connected via 50,828 edges. Dataset summary statistics for the QS top 50 is given in Table 1. For the sake of brevity, only the top 10 of top 50 ranked institutions from the 2016 QS subject area 'dentistry' is given in . A summary of selected *overall* dentistry network statistics for QS top 10 is given in **Error! Reference source not found.**.

As mentioned in the methods section the 'ego' network was analyzed for each of the QS top 50 institutions. The number of nodes and edges in each ego networks were noted and the resultant metrics are given in Table 4 (only for the QS top 10).

Table 1 - QS top 50 dataset summary statistics

Dataset	Average	Median	StDev	Max	Min
Document count	159	120	119.6	706	42
Citation	8.67	8.76	2.19	13.26	4.61
RCI_{year}	1.36	1.38	0.33	2.01	0.73

Table 2 - 2016 QS top 10 ranked dentistry institutions

QS Rank	Institution
1	University of Hong Kong
2	University of Michigan
3	Karolinska Institutet
4	King's College London
5	Goteborgs Universitet
6	Tokyo Medical and Dental University
7	KU Leuven
8	University College London
9	Universidade de Sao Paulo - USP
10	New York University

Table 3 - 2016 QS data for top 10

Affiliation	**QS Rank**	**Overall Score**	**Academic Reputation**	**Employer Reputation**	**CPP**	*Dataset Doc count*
University of Hong Kong	1	91.5	80.2	97.2	96.8	*173*
University of Michigan	2	88.6	71.1	84.5	96.1	*318*
Karolinska Institutet	3	86.7	85.8	94	88.5	*113*
King's College London	4	86.4	83.1	69.7	89.2	*203*
Goteborgs Universitet	5	85.8	74.8	58.6	95.6	*91*
Tokyo Medical and Dental University	6	85.5	100	83.3	80.1	*175*
KU Leuven	7	84.9	71.8	62.1	100	*129*
University College London	8	84.3	65.9	82.1	91.5	*239*
Universidade de Sao Paulo - USP	9	83.4	64.6	96.3	86.2	*706*
New York University	10	83.3	69	74.4	90.5	*211*

Table 4 - 'Ego' network data and dataset citations for the QS top 10

Affiliation	**'ego' Nodes**	**'ego' Edges**	**Edges/Node**	**Dataset Average Citations**	**Dataset Average RCI_{year}**
University of Hong Kong	138	2101	15.22	9.66	1.48
University of Michigan	317	5799	18.29	8.81	1.52
Karolinska Institutet	119	1630	13.69	11.27	1.61
King's College London	194	3005	15.49	7.65	1.07
Goteborgs Universitet	134	3004	22.42	12.15	1.87
Tokyo Medical and Dental University	119	1433	12.04	4.76	0.80
KU Leuven	123	2239	18.20	10.68	1.86
University College London	266	4584	17.23	8.72	1.29
Universidade de Sao Paulo - USP	354	5618	15.87	4.81	0.73
New York University	296	5718	19.32	7.38	1.19

There was a weak but positive correlation (Pearson's 'r' 0.3375, P-value 0.016534, where N=50) between the dataset institutional citation average and the ratio of 'ego' networks Edges/Node. This result is significant at $p<0.05$. Moreover, the Spearman's correlation was calculated and the value of R = 0.338583 and the two-tailed value of P = 0.01617 was obtained. This result between the two variables would be considered statistically significant.
The same two correlations were performed between institutions' dataset RCI_{year} and the 'ego' networks links/node ratio. Pearsons 'r' = 0.3692, P-value 0.0083 which is significant at $p<0.05$. Spearman R = 0.3863 and the two-tailed value of P is 0.00559. This association between the two variables would again be considered statistically significant.

Other studies have also shown relationships between network characteristics and citation data [8, 15-17]. A study by Billah and Gauch [18] reported the prediction of success for young researchers using social network analysis. Their definition of "success" seemed to be related to h-index which they acknowledged is low in researchers early in their career. What is interesting in Billah and Gauch's study is that they claim triumph in predicting a researcher's success using an analysis of their professional network yet they seem to miss a very important point raised by their results. If a researcher's professional network is important for predicting their future "research impact", and if the dynamics of their 'neighborhood' has strong positive impact on an author's prospect in the future then perhaps it is not the "quality" of one's research that drives citations but the connections you have? Indeed, this seemed to be the premise of Billah and Gauch's research when they write their hypothesis that "young researchers with strong social connections to established researchers are more likely to have successful research careers". If the measure of success is h-index, again, the question is what drives what?
Lastly, the University of Hong Kong (UHK) was ranked #1 in the 2016 QS dentistry ranking whereas the University of Bristol (Bristol) was ranked #50. Combining their 'ego' co-authorship networks gives Figure 2; the UHK node and its edges are colored pink and Bristol's node and edges are blue (node size is proportional to degree). This image clearly shows UHK has a larger and more comprehensive co-authorship network compared to Bristol – 15.22 edges per node compared to 8.84. Given the network vs

citation correlations described in this paper and other reported studies, one could suggest that UHK's high rankings is a consequence of its networks rather than a consequence of "quality". Continuing success therefore would seem to be a combination of changing network dynamics and the 'Matthew Effect' associated with a particular institution. Perhaps this is one explanation why, when it comes to university ranking systems, the mix of "top ranked" universities has remained virtually stagnant over the years.

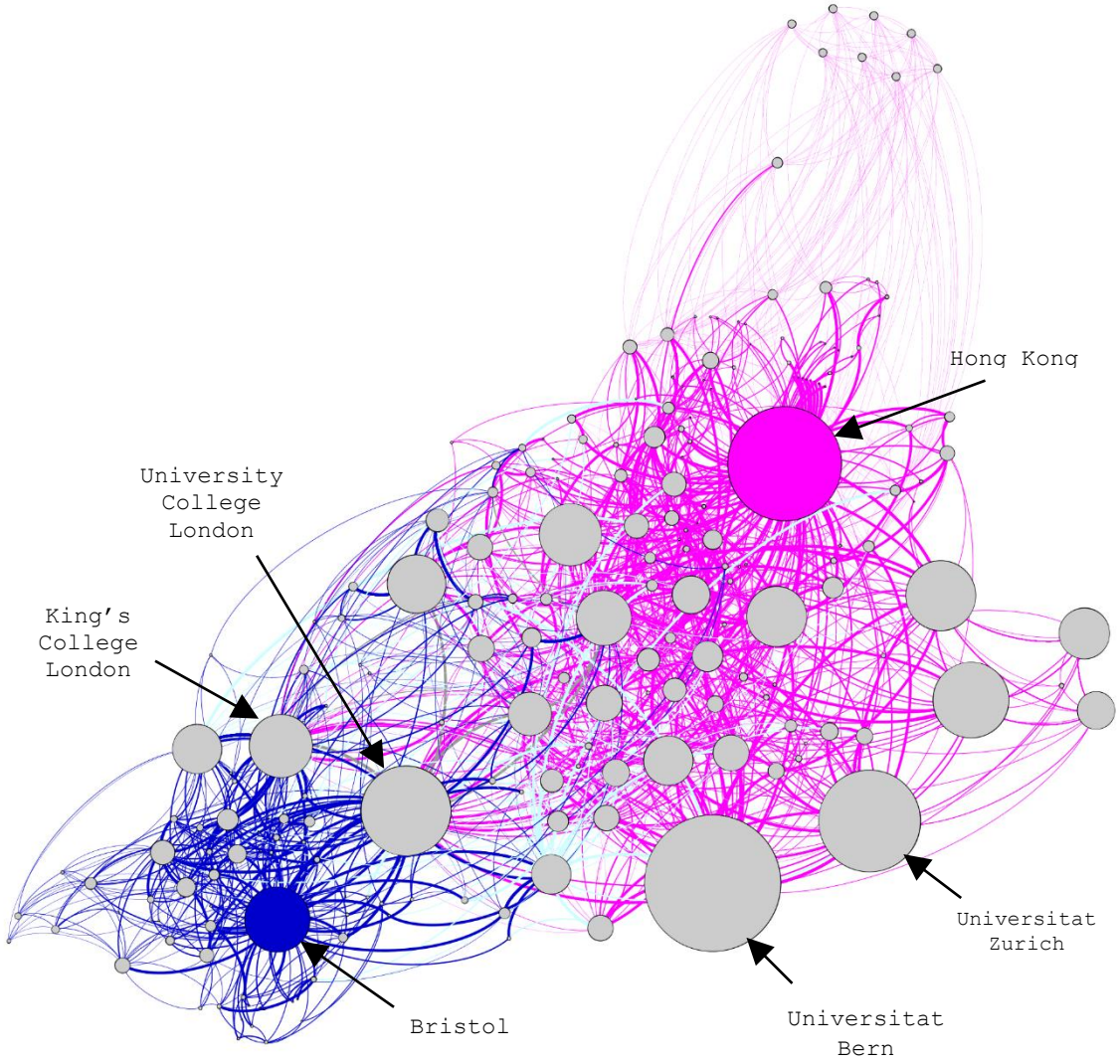

Figure 2 - University of Hong Kong 'ego' network (138 nodes, 2101 edges)

Conclusion

Are citations the result of "quality" or "excellence" as is commonly accepted or are citations driven by developing the right connections? This paper and others like it seem to suggest there is an increasing argument for citation being driven by the right connections. Thus, citation counting has have very little to do with "quality", "excellence" or "impact".

With the on-going development of social network analysis and the evolution of thought that networks provide powerful insights into human behavior, perhaps it is time for

the academic community (and its administrators) to acknowledge the importance of 'not what you know but who you know'. Perhaps it is time to begin the process of severing the umbilical cord that seems to exist between "quality" and citation counting.

References

1. Hirsch, J.E., *An Index to Quantify an Individual's Scientific Research Output.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(46): p. 16569-16572.
2. Kosmulski, M., *Hirsch-type index of international recognition.* Journal of Informetrics, 2010. **4**(3): p. 351-357.
3. Jin, B., et al., *The R- and AR-indices: Complementing the h-index.* Chinese Science Bulletin, 2007. **52**(6): p. 855-863.
4. Garfield, E., *Citation Indexes for Science.* Science, 1955. **122**(3159): p. 108-111.
5. Garfield, E., *Is citation analysis a legitimate evaluation tool?* Scientometrics, 1979. **1**(4): p. 359-375.
6. Wade, N., *Citation Analysis: A New Tool for Science Administrators.* Science, 1975. **188**(4187): p. 429-432.
7. Newman, M.E., *Coauthorship networks and patterns of scientific collaboration.* Proc Natl Acad Sci U S A, 2004. **101 Suppl 1**: p. 5200-5.
8. Abbasi, A., K.S.K. Chung, and L. Hossain, *Egocentric analysis of co-authorship network structure, position and performance.* Information Processing and Management, 2012. **48**(4): p. 671-679.
9. Biscaro, C. and C. Giupponi, *Co-authorship and bibliographic coupling network effects on citations.* PLoS ONE, 2014. **9**(6).
10. Merton, R.K., *The Matthew Effect in Science: The reward and communication systems of science are considered.* Science, 1968. **159**(3810): p. 56-63.
11. Bonitz, M., E. Bruckner, and A. Scharnhorst, *Characteristics and impact of the matthew effect for countries.* Scientometrics, 1997. **40**(3): p. 407-422.
12. Bonitz, M., E. Bruckner, and A. Scharnhorst, *The matthew index—Concentration patterns and Matthew core journals.* Scientometrics, 1999. **44**(3): p. 361-378.
13. Bonitz, M., *Ten years Matthew effect for countries.* Scientometrics, 2005. **64**(3): p. 375-379.
14. Kern, R., M. Zechner, and M. Granitzer. *Model selection strategies for author disambiguation*. 2011.
15. Bertsimas, D., et al. *Network analysis for predicting academic impact*. 2013.
16. Uddin, S., L. Hossain, and K. Rasmussen, *Network Effects on Scientific Collaborations.* PLoS ONE, 2013. **8**(2).
17. Newman, M.E.J., *The structure of scientific collaboration networks.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(2): p. 404-409.
18. Billah, S.M. and S. Gauch. *Social network analysis for predicting emerging researchers*. in *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2015*. 2015. SciTePress.