Identifying Research Talent Using Web-Centric Databases

Anca Dumitrache[1]   **Paul Groth[2]**   **Peter van den Besselaar[3]**

Network Institute, VU University Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam (The Netherlands)
[1]anca.dumitrache@student.vu.nl, [2]p.t.groth@vu.nl, [3]p.a.a.vanden.besselaar@vu.nl

ABSTRACT
Metrics play a key part in the assessment of scholars. These metrics are primarily computed using bibliometric data collected in offline procedures. In this work, we compare the usage of a publication database based on a Web crawl and a traditional publication database for computing scholarly metrics. We focus on metrics that determine the independence of researchers from their supervisor, which are used to assess the growth of young researchers. We describe two types of graphs that can be constructed from online data: the co-author network of the young researcher, and the combined topic network of the young researcher and their supervisor, together with a series of network properties that describe these graphs. Finally, we show that, for the purpose of discovering emerging talent, dynamic online publication resources provide better coverage than more traditional datasets, and more importantly, lead to very different results.

Author Keywords
Scholarly networks, bibliometrics, altmetrics, online vs. offline databases, independence indicators

ACM Classification Keywords
J.4 SOCIAL AND BEHAVIORAL SCIENCES

INTRODUCTION
The availability of sources of publication data including citations has changed the way the assessment of scholars and scholarship is performed and has impacted our understanding of the science system. Bibliometrics now play a central role in how we judge science.

Traditionally, the primary source of data about publications and researchers has been Thomason Reuters' Web of Science. This is a curated database of publications produced using data provided by various publishers who have been selected for inclusion in the database. While the data within the Web of Science is relatively clean, it suffers from a lack of coverage, as many scholarly venues are not covered [6].

However, with the rise in popularity of online sources of scholarly information, there is increasing interest in their use for bibliometrics and other forms of science studies [5, 7]. Examples of such sources include Mendeley, where researchers can list their own publications, and Google Scholar, which crawls the web for scientific publications and compiles lists based on author names.

Here, we investigate the usage of data from such an online source (Google Scholar) for the calculation of independence metrics. These metrics are used to analyze how talent develops in scholarly networks. The indicators describe the relation between a young researcher and their mentor (typically their PhD supervisor). Independence in this case refers to the ability of the young researcher to explore different research directions other than their (former) supervisor.

We compare the results of applying these metrics on data from Google Scholar[1] with the results from their application to Web of Science[2] data for two cases from social science and computer science. Our results indicate that web-based datasets have both better coverage and are faster in the inclusion on new information. The inclusion of recent data is essential for determining independence, as young researchers have had less time to build up a record of work. Furthermore, the wider coverage may result in a different (and better) picture.

The rest of this paper is organized as follows. We begin with discussing of related work followed by a description of our approach for measuring independence. We then present results and conclude.

RELATED WORK
Kurtz et al. [5] reviewed how metrics for online data perform against bibliometrics. Given the difficulty in extraction of citation information from online sources, the authors noted the need for new metrics that are tailored to the available data. Priem et al. [7] looked at metrics for evaluating publications retrieved online. Their main focus was on community-generated content such as blogs. They argue that analyzing online data can provide a new outlook on scholar interconnection by adding a dynamic

---

[1] http://scholar.google.com

[2] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

factor – the users generate content, thus enriching the information at a faster pace than a standard database would be able to. Other studies have shown that typical bibliometric measures such as productivity and citation impact, are a poor predictor for success grant applications [4].

The lack of valid performance indicators at the individual level led to the development of independence metrics [8] that may better predict success in academic careers. For an emerging researcher, independence is defined in relation to the former supervisor. Van den Besselaar introduced two indicators that can be deployed for identifying talent: (i) an indicator measuring the *own* co-author network of the young researcher, independent and (ii) an indicator measuring the *own* research lines of the young researcher compared to the research of the former supervisor. These metrics were tested using data from a curated database of scholarly publications.

APPROACH

For the purpose of this experiment, we selected a set of young researcher and supervisor pairings, for which all the necessary data was retrieved from Google Scholar. For each pair, the data was organized into two networks: the co-author network of the young researcher, and the combined paper network of the young researcher and their supervisor. The indicators used for determining researcher independence were adapted from the research by Van den Besselaar et al. [8]. However, for describing the similarity in research topics we did not use citations, but title similarity, which are better suited to online datasets (as citations are not available).

Indicator 1: The structure of the co-author network

The co-author network of the young researcher refers to the network of authors with whom the young researcher has published papers. It was constructed by retrieving the publication list of the young researcher. Every author corresponds to a node in the graph, whereas edges represent collaborations on the same paper. The young researcher is at the center of the network, and has edges with every other node in the network, as the graph was constructed based on their publication list. Edges are unweighted. Two properties of this network are used for measuring independence.

The *eigenvector centrality of the supervisor* [3] in the researchers' co-author network indicates the number of collaborations between the young researcher and their supervisor in comparison to their other collaborations. The higher the independence of the young researcher, the more of his/her coauthors are no coauthors of the supervisor. This leads to a lower eigenvector centrality (closer to zero) of the supervisor. In contrast, the centrality of the young researcher in their own network is always equal to 1.

The clustering coefficient of the young researcher in his/her own network is always low (close to 0). The *clustering coefficient of the supervisor* [9] in the young researchers' co-author network indicates the extent to which the supervisor has the same circle of coauthors. Consequently, the higher the clustering coefficient of the supervisor, the more the supervisor is located in a smaller part of the researchers' coauthor network, and the more independent the researcher is.

Indicator 2: The topic network of the researcher

The topic network of the researcher refers to the combined networks of the scientific papers authored by the young researcher and the former supervisor. It was constructed by retrieving the publication list of both researchers, and combining them. Each node in this network refers to a paper.

In the original version by Van den Besselaar, a topics network was constructed based on the bibliographic coupling between papers. However, because metadata on references in papers is not readily available on the Web, we considered a different metric for constructing this network.

Edges were added based on the similarity between paper titles. Each title was processed so as to reduce the number of false positive results, by punctuation removal, stemming, and stopword removal. Then, for each pair of papers, the number of words in common was computed, and then normalized over the length of the title with the least number of the words. Results with a similarity coefficient smaller than 0.2 were discarded, and a weighted graph was constructed. Isolated nodes were also discarded.

The network of paper title similarity can be used to study how the topics of both the young researcher and their supervisor overlap. By looking at the clusters in the network, one can identify lines of research. Independence would be indicated by the existence of clusters of papers authored by the young researcher only, but also by the number of isolated papers of the young researchers.

RESULTS

The independence indicators were applied to a set of researcher-supervisor pairs: one from computer science (researcher R1, supervisor S1), and one from the social sciences (researcher R2, and supervisor S2).

The co-author network

Figure 1 illustrates the co-author network of R1, where the nodes corresponding to both R1 and S1 are highlighted. S1 appears in a cluster with a significant number of other co-authors (left side of the image). However, other smaller clusters (center and right side) are also apparent. This seems to indicate that, while a sizable amount of the collaborations were performed inside the cluster of the supervisor, the young researcher has also branched out into a series of separate collaborations.
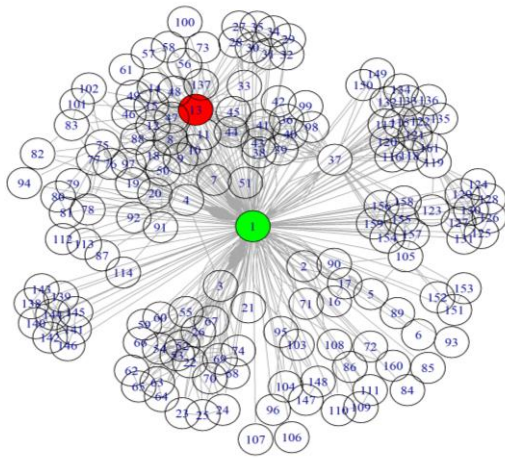
Figure 1. Co-author network for the selected case, with R1 represented as node 1 in green, and S1 as node 13 in red.
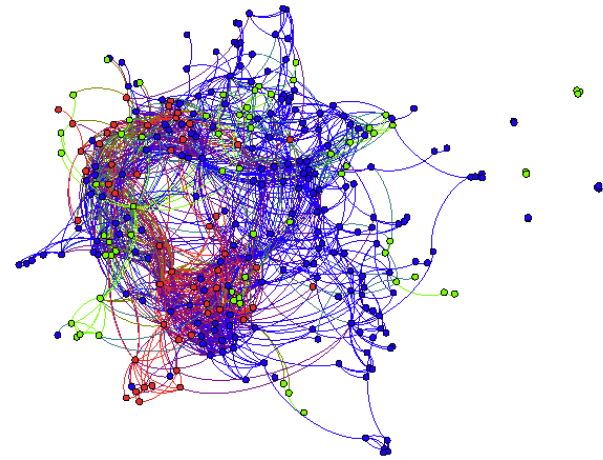

Figure 2. The combined topic network of young researcher R1 and the former supervisor S1, containing publications authored only by R1 (green nodes), only by S1 (blue nodes), and joint publications (red nodes).

Network property	R1	R2
Papers retrieved	128	29
Share of papers with former supervisor	41.4%	35.9%
Number of co-authors	160	47
Median number of relations with a co-author	3.2	2.6
Eigenvector centrality of young researcher	1	1
Eigenvector centrality of supervisor	0.12	0.64
Clustering coefficient of young researcher	0.05	0.13
Clustering coefficient of supervisor	0.18	0.43

Table 1. Property values for the co-author network

The values of the network properties, shown in Table 1, confirm the results of the visual inspection of the graph. R1 has a large share of common papers with their former supervisor S1, and the clustering coefficient of S1 is quite low, indicating that S1 has indeed collaborated with many of the co-authors of R1. The eigenvector centrality of S1 is quite low as well. This seems to indicate that the supervisor is not very central in the researchers' network, meaning that there exist quite some collaborations of R1 that are independent from their former supervisor.

Another reason why the eigenvector centrality of S1 has such a low score is that the co-author network is unweighted, meaning that a link is only counted once, even if it represents more coauthored papers. As R1 shares a large number of papers with S1, the value for the eigenvector centrality for S1 is significantly higher in the weighted co-author graph. By analyzing the unweighted graph of co-authors, we ensure that we can detect new

directions in the work of the young researcher, without having the many early career collaborations with their former supervisor cloud the results.

The topic network

Figure 2 shows an illustration of the combined topic network for R1 and S1. The coloring of the nodes in the graph serves to identify three classes of papers: (1) publications only authored by R1, (2) publications only authored by S1, (3) publications authored by both R1 and S1. Publications in class 2 dominate in the network, as the dataset retrieved for supervisor S1 was significantly larger than for researcher R1. Using Gephi[3] software for graph analysis, a clustering algorithm [1] was performed on the network, in order to identify the different lines of research as clusters publications. 16 clusters were computed for the combined network of R1 and S1.

The topic network consists of a large subgraph of interconnected publications, containing several clusters. As expected, most of the topics are dominated by papers of S1, and a few by joint work of S1 and R1. However, in this paper we are interested in whether researcher R1 also started to work on research topics different from the research interests of his/her former supervisor S1. Inspecting the graph, we indeed find at several places small concentrations of green nodes, only weakly linked with the blue or red nodes. These green clusters indicate emergent independent research of R1.

The topic network also contains several small isolated clusters, consisting of publications from the same class. These clusters have a high degree of interconnectivity, indicating standalone directions in academic research. Finally, it is worth noting that 13 publications authored by R1 (10% of the dataset) did not appear in the network at

---

[3] https://gephi.org/

all, as the title similarity between them and other titles was negligible. This indicates that, in addition to the clusters already discussed, there exist a series of isolated topics that R1 has published on. These results seem to indicate that the researcher has begun branching into different directions from their supervisor.

Comparison with a standard database
In contrast with Google Scholar, which crawls the Web for data on published papers, the Web of Science provides a curated database of publications. The latter is less quickly updated and has a lower coverage than the former (and other community-generated sources). We now compare the independence indicators based on the two sources: GS and WoS.

Table 2 shows that there are fewer publications available for both researchers in the Web of Science, as compared to Google Scholar. This is of course expected. However, the question is whether different coverage also has implications for the measured independence.

Network property	R1	R2
Papers retrieved	24	8
Share of papers with former supervisor	58%	38%
Number of co-authors	109	8
Median number of relations with a co-author	6.9	1.4
Eigenvector centrality of young researcher	1	1
Eigenvector centrality of supervisor	0.94	0.63
Clustering coefficient of young researcher	0.26	0.14
Clustering coefficient of supervisor	0.61	0.67

Table 2. Property values for the co-author retrieved from the Web of Science.

As table 2 shows, the values for the network indicators are also different. In the case of R1, for whom only 24 papers were retrieved from the WoS, the influence of the supervisor in the WoS-network is large, as evidenced by the high value of the eigenvector centrality. This in contrast with the metrics for the online data, which highlighted the independence of R1.

The difference was not as drastic in the case of R2, where the same eigenvector centrality was computed for both datasets. However, value of the clustering coefficient is larger in this dataset, again showing that R2 appears to be more independent in the online dataset.

CONCLUSION AND FUTURE WORK
In this work, we compared the use of online and offline sources for determining scholarly independence. We demonstrated that aggregated content sources such as

Google Scholar not only provide a wider scope, as they factor in more diverse research output and are updated at a faster pace. Moe importantly, traditional bibliometric databases obviously are not representative for the larger output covered by online data, and this may result in very different (relative) performance scores, as we showed for researcher R1.This makes online sources more suitable when assessing early career researchers.

In the future, we will expand our study to larger samples of supervisee-supervisor relations, using more online data sources, and different indicators for topic similarity beyond the scope of lexical similarity. Finally, our larger aim is to discover trends in how scholarly talent develops, using the Web (e.g. after how many years does a researcher usually need to discover their niche).

REFERENCES
1. Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

2. Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6), e6022.

3. Bonacich, P. (1987). Power and Centrality: A Family of Measures. American Journal of Sociology, 92, 1170-1182.

4. Bornmann, L., Leydesdorff, L., and Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: different ways of comparing rejected to awarded applications. *Journal of Informetrics*, 4(3), 211-220.

5. Kurtz, M. J., and Bollen, J. (2010). Usage bibliometrics. *Annual review of information science and technology*, 44(1), 1-64.

6. Norris, M. and Oppenheim, C. Comparing alternatives to the Web of Science for coverage of the social sciences literature. *Journal of Informetrics*, vol. 1, no. 2, pp. 161–169, 2007.

7. Priem, J., and Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7).

8. Van den Besselaar, P., Sandström, U., and Van der Weijden, I. (2012). The Independence Indicator: Towards Bibliometric Quality Indicators at the Individual Level.

9. Wasserman, S., and Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.