Titre: Title:	The study of network effects on research impact in Africa
Auteurs: Authors:	Leila Tahmooresnejad, Catherine Beaudry, & Seyed Reza Mirnezami
Date:	2021
Type:	Article de revue / Article
Référence: Citation:	Tahmooresnejad, L., Beaudry, C., & Mirnezami, S. R. (2021). The study of network effects on research impact in Africa. Science and Public Policy, 48(4), 462-473. https://doi.org/10.1093/scipol/scab030

Titre de la revue: Journal Title:	Science and Public Policy (vol. 48, no. 4)
Maison d'édition: Publisher:	Oxford Academic
URL officiel: Official URL:	https://doi.org/10.1093/scipol/scab030
Mention légale: Legal notice:	

The Study of Network Effects on Research Impact in Africa

Leila Tahmooresnejad[1*], Catherine Beaudry[2], Seyed Reza Mirnezami[3]

[1,2]Polytechnique Montreal, Montreal, QC, H3C 3A7, Canada, [3]SPRI, Sharif University of Technology, Tehran, Iran

Abstract

This paper studies the relationship between the position of individual scientists within co-authorship networks and their scientific performance. Using co-authorship data from African scientists in the Health and Medical Sciences within a timespan of fifteen years (2000-2015), we characterize the collaboration networks and calculate centrality measures for each scientist to explore how scientific production and impact can be associated with their position within the network. Our findings reveal that authors who occupy a better position within their network and are deemed to actively collaborate with others also have a higher research impact. In this regard, South African scientists do not differ from those in the rest of the world.

Keywords: Co-authorship, Social network analysis, Research impact, Africa

Introduction

According to Pouris and Ho (2014), single-author articles seem to be disappearing in Africa due to the effect of foreign funding sources. A rising awareness regarding the importance of strong research links within the scientific network has led to a sharp increase in the number of collaborations. As in the rest of the world, African scientists are embedded in research networks that influence and are influenced by their collaborations with other scientists and the recognition they receive therein. The complex set of interactions between co-authors influence the number of papers published and the number of citations they receive.

In a networked environment, the position of each scientist plays an important role in diffusing knowledge. According to the literature, the number of citations is the most important indicator for

[1] leila.tahmooresnejad@polymtl.ca
[2] catherine.beaudry@polymtl.ca , Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Tel: +1 514 340 4711 x3357

[*] Corresponding Author: leila.tahmooresnejad @polymtl.ca, Department of Mathematics and Industrial Engineering, Polytechnique Montreal, P.O. Box. 6079, Downtown office, Montreal, QC, H3C 3A7, Canada. Tel.: +1 514 340 4711; fax: +1 514 340 4173.
[3] srmirnezami@sharif.edu, Sharif University of Technology, Tehran, Iran

measuring the ‘success’ of scientific research. While the relationships between scientists may affect the visibility of their research, it is not still clear how the citation counts are related to network positions, especially in Africa. Studying networks of African scientists will help broaden our understanding of the collaboration relative to scientific performance measures. Some attempts have been made to find associations between network measures and citation counts (Otte and Rousseau, 2002; Tahmooresnejad and Beaudry, 2018; Yan and Ding, 2009), but this relationship is not yet well-understood, and has never been examined in Africa.

The use of social network analysis has grown in importance over the recent years, particularly for identifying influential scientists. Co-authorship links offer a great opportunity to investigate the structure of scientific collaborations. As the number of co-authors grows with time, so does the size of co-authorship networks, making their structures even more complex. It is therefore a requirement to find the most appropriate measures to assess such structures and understand the collaboration patterns. The study of these complex network structures provides a rough indication of how knowledge flows and is exchanged among scientists prior to publishing articles. The sharing of ideas and experiences within scientific networks have gained the attention of scholars interested in detecting the patterns of such collaborations.

This paper contributes to advancing knowledge on the use of social network analysis in measuring the role of scientific collaborations in the African publication landscape. African scientific co-authorship has become an interesting topic as researchers try to shed light on the dynamics and motives for collaboration in the African research system. Pouris and Ho (2014) pointed out that collaboration patterns within the African continent are substantially higher in comparison to the rest of the world. They also found that during the five-year period of 2007-2011, the share of international collaborations grew by 66%.

In this study, we characterize the co-authorship network of authors who published articles between 2000-2015. We investigate the relationship between network measures and the performance of authors based on scientific production (publication counts) and impact (citation counts). This paper adds to the very few studies on the impact of collaborations and research output in Africa and thereby contributes to the current debate on the importance of collaborations for scientists in Africa, and particularly on how the internal and international collaborations among various countries and institutes affect research performance in Africa.

To perform this analysis, we build dynamic five-year networks to measure the evolution of collaboration over the years. We construct a variety of measures to investigate the relationship between network measures and the research output of scientists. We perform classic Social Network Analysis (SNA) to calculate various centrality indicators and clustering coefficient measures to identify which are best related to the performance of African scientists in the Health and Medical Sciences which includes three domains: Health, Biomedical Research and Clinical Medicine.

Our research fills a gap in the literature regarding the influence of collaboration on scientific production and impact using detailed network measures of co-authorship for African scientists. The contribution of this paper is four-fold. First, we show that better connected South African health scientists, i.e. those that occupy more central positions in the co-authorship network, generally contribute to a larger number of publications which are then more cited. Second, diversifying the number of unconnected groups of researchers with whom they collaborate is also associated with greater scientific production and impact. In essence, South African scientists follow the pattern observed in the rest of the world. Third, while collaborating internationally more intensively will bring an increased number of citations for these scientists, it has no effect on the number of papers published, in other words, it increases quality, not quantity. Fourth, co-authoring papers with fellow Africans and South-Africans both contribute to increasing scientific production and impact. As a consequence, African science policy should not favor international collaboration (outside of the African continent) to the detriment of intra-African collaboration. Intra-Africa and international mobility programs should be further expanded, to mitigate the costs of such programs.

The remainder of the paper is organized as follows. In the Background and related work section, we review the existing studies on co-authorship networks and how the performance of researchers is associated with different Social Network Analysis measures. The Methodology section then presents the dataset, the network measures and the regression models. A brief descriptive analysis is followed by the Results section which discusses the correlation between various network measures and research performance (scientific productivity and impact) using conventional statistical analysis. Finally, we conclude the paper with a discussion on the results and propose a few suggestions for future lines of investigation.

Background and related work

There is a constant interest in the role of collaboration within the academic realm. Studies of collaboration have generally considered two approaches to measure the phenomenon: formal and informal collaboration among scientists. The former is assessed via joint papers and participation in various research groups. The latter arises from joint work involving teacher-student relationships, conference attendance and related networking, reviewing, journal editing, discussions with colleagues, etc.

It is not clear which of these measurement approaches is the most relevant to evaluate collaboration within the academic community. Many studies have nonetheless focused on co-authorship as one of the indicators of formal collaboration since it is the most tangible, easily accessible, measurable, and well-documented forms of collaboration. Formal collaboration is basically easier to analyze in comparison to informal collaboration. The next section reviews the pros and cons of using such measures of collaboration.

Co-authorship networks

According to Laudel (2002), co-authorship is not an adequate measure of collaboration because all individuals who are involved in the preparation of a paper are not cited as co-authors. In fact, co-authors are a subset of the contributors who are acknowledged as authors of the paper. Katz and Martin (1997) argued that there is a positive correlation between collaboration and co-authorship, but that it can only be considered as a partial indicator of collaboration. However, Glänzel and Schubert (2004) highlighted that studying the scientific collaborations through co-authorship provides a deep insight into the interaction between collaboration, scientific communication, and the performance of scientists.

Collaboration between scientists is also a complex phenomenon which is influenced by both economic and scientific factors. For example, research funding, scientists' mobility, or any change in their communication patterns, all affect collaboration. Even papers with a single author who is associated with multiple institutions implicitly manifest a form of collaboration by the very involvement of these institutions and countries between which the author is mobile (Glänzel, 2001). Glänzel and Schubert (2004) explored a higher level of aggregation in considering institutional collaborations. They found that inter-institutional collaborations can uncover an interesting phenomenon regarding the effect of multiple affiliations, and thus association between different institutions, on collaborative patterns.

Co-authorship networks have been studied extensively to understand the structure of scientific collaborations. In contrast with co-citation links, co-authorship affords a much stronger connection because co-authors generally know each other and have collegial relationships. Citations, on the other hand occur over a period of time and without any knowledge or relationship with the person or article cited.

Not only does collaboration foster higher productivity, it also influences the citation process. Citing others and receiving citations has been shown to differ for co-authored papers as opposed to single-authored papers. Persson et al. (2004) found that the list of references is longer in co-authored papers, largely because each author adds references to the list. As a consequence, the average number of references grows in direct relation with the number of authors.

Co-authorship collaborations and social networks

Similarly to the formal and informal measures mentioned above, in recent decades, different approaches have been considered for studying co-authorships. These approaches explore the reasons for scientists to collaborate and tries to analyze the consequences of these collaborations.

The structure of co-authorship networks has been studied by Newman (2001, 2003, 2004) in order to understand how these collaboration patterns vary over time. The analysis of such structures provides insight into the structural changes related to collaboration. Co-authorships have proved useful to build the social network for testing various related theories and explore the structural properties of large academic networks. The questions that arise for our research are: 1) what social structure can be identified in these networks? and 2) how can this structure affect the impact of researchers?

Social network analysis is essentially based on the relationship between actors. The graph's nodes represent actors and the edges represent the interactions between nodes. In co-authorship networks, authors are represented by nodes which are connected if individuals have co-authored one or more papers. Characterizing academic networks using social network graphs at the individual level allow scholars to use graph theory to analyze the position or status of each actor in the network. Their position is usually expressed in term of centrality, i.e. measuring how central is the author in the network. Centrality is one of the most widespread concepts in social network analysis. As its name suggests, centrality refers to the position (from the periphery to the center) of an author and how prominent is his or her role in the network (Troshani and Doolin, 2007).

There are three typical measures of centrality in the social network literature: degree centrality, closeness centrality and betweenness centrality. Degree centrality is defined as the number of direct connections (co-authorship links) that a given author shares with other authors. An author with a higher degree centrality has therefore collaborated with many other authors in publishing, most probably, a large number of papers (Tahmooresnejad and Beaudry, 2018).

Closeness centrality of an author expands the definition of degree centrality and determines how close an author is to the others in the network. An author with a higher closeness centrality has many short connections to the other authors. In essence, this measure is the inverse attribute of degree centrality, which means that a more central author has a smaller value of closeness centrality. In this regard, it is defined as the mean shortest distance by which an author is connected to all other authors in a graph. An author with the highest closeness centrality can reach others with the minimum number of intermediary nodes (Otte and Rousseau 2002). Borgatti (2005) interpreted closeness as a measure of the expected time necessary to receive communications through the network.

Betweenness centrality is another well-known centrality measure defined as the number of times (in proportion) that a given author is "involved" in connecting two other authors via the shortest path across the network. More simply explained, it is the number of times that an author is found on the shortest path between any two authors in the network. The concept considers the proportion of the shortest paths between two authors that would pass through a given author, who is then deemed an important intermediary. Betweenness centrality measures the extent to which an author can facilitate the flow of knowledge, resources, or information within the network. Thus, an author with a high level of betweenness centrality can be considered as a crucial bridge to control the flow of knowledge across different parts of the network (Freeman, 1978; Boragatti, 2005; Li et al., 2013).

Closely associated with the notion of centrality is the concept of "cliquishness" in the network. Past studies have considered the clustering coefficient or "cliquishness" to measure the tendency of authors to cluster together. A clique in the network is a subnetwork in which any author is directly linked to any other author of the subnetwork. This measure represents the likelihood that author A and author B are connected if author A and author C have a relationship and there exists a relationship between author B and author C as well. The value of cliquishness is always a number between 0 and 1. In order to have a better understanding of this indicator, we present three simple examples of networks with different measures of cliquishness for author A: the cliquishness for author A is higher in Figure 1c (equal to 1) and lower in Figure 1a (equal to 0).

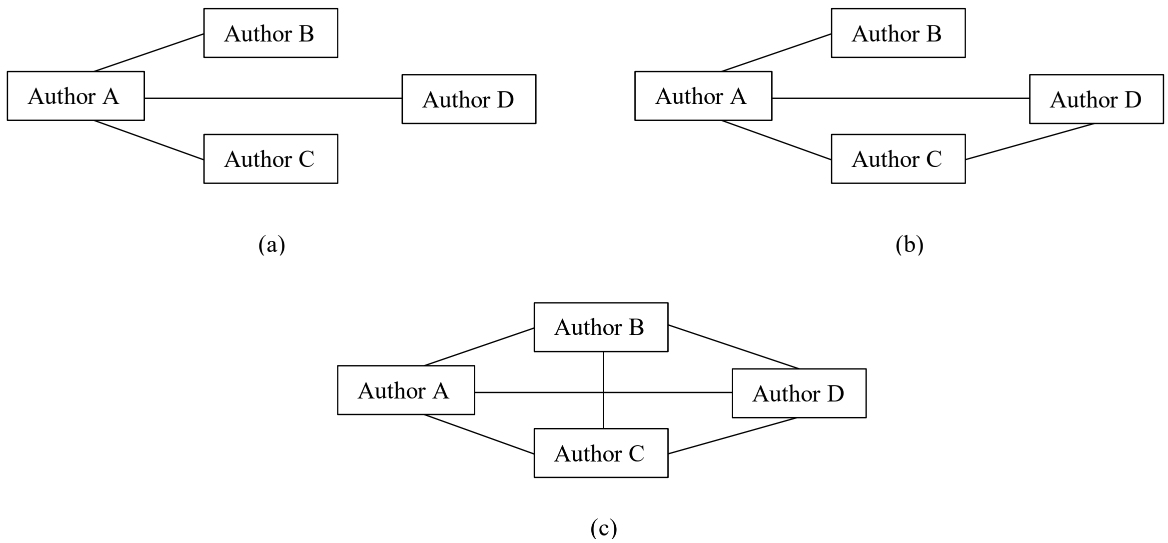

Figure 1- Example of three networks with different cliquishness

The impact of co-authorship collaborations on research

The empirical literature has examined whether co-authorship links lead to the higher productivity of scientists. Hamermesh and Oster (2002) pointed out that some authors may choose to work with other individuals with the same skill level. Two authors, however, are more likely to collaborate if they expect their joint productivity from this collaboration to be higher compared to their productivity from solo work. Because collaborations potentially increase access to new ideas and resources, Lee and Bozeman (2005) highlighted that the positive impact of co-authorship on productivity depends on the authors with whom individuals choose to collaborate.

Lindenlaub and Prummer (2020) found that while degree centralities are beneficial for the performance, clustering plays a peer pressure role and its impact depends on the environment of the individuals and the gender. In another study by Ductor et al., (2018), high degree and low clustering were shown to be correlated with higher research performance.

In line with the research findings of prior studies, we propose that having a central position in co-authorship networks increases an author's chances of connecting to prolific scientists and is therefore more likely to gain a shared expertise which enhances his or her research impact. Our first hypothesis therefore reads as follows:

H1: Authors that occupy a more prominent position in a co-authorship network also publish a greater number of articles

H1a: *Authors with higher degree centrality publish a greater number of articles.*

H1b: *Authors with higher betweenness centrality publish a greater number of articles*

H1c: *Authors with higher closeness centrality publish a greater number of articles*

H1d: *Authors with higher clustering coefficient publish <u>fewer</u> articles.*

Chung et al., (2009) found that the number of citations received by papers that are written by a larger team of coauthors is greater than for single-authored papers. Prior studies have illustrated that the research impact may be increased by co-authorship links. For example, Yan and Ding (2009) found that there is a positive correlation between network measures and citation counts in co-authorship networks.

Abbasi et al. (2011) and Li et al. (2013) studied the impact of centrality on research performance. Their findings revealed a significant and positive effect of degree centrality on research performance. They also highlighted that because an author with a high degree centrality is connected to many authors, that author would therefore benefit from the information and knowledge flow between these other authors. It is this connection that increases the likelihood that these benefits will improve the quality of papers, and consequently that these papers should receive more citations. Another possible explanation stems from the fact that his or her work is more likely to be seen by the authors who are directly or indirectly linked to that highly central individual. Hence, we expect that central positions in a co-authorship network would lead to higher numbers of publications and citation counts.

Because prolific scientists are also prominent in their network, we postulate that:

H2: Authors that occupy a more prominent position in a co-authorship network are more cited.

H2a: *Authors with higher degree centrality are more cited.*

H2b: *Authors with higher betweenness centrality are more cited.*

H2c: *Authors with higher closeness centrality are more cited.*

H2d: *Authors with higher clustering coefficient are <u>less</u> cited.*

Methodology

Data

The data used in this study have been extracted from the Leiden[4] database of the Web of Science, from which we have extracted all the articles related to the Health and Medical Sciences for the period 2000-2015. In order to identify the African scientists, we extracted all publications that had at least one African affiliation. Our data consists in a panel of African scientists that cover a fifteen-year time span. The dataset contains three domains in Health and Medical Sciences: Health, Clinical Medicine and Biomedical Research.

Dependent variables

Our analysis considers two dependent variables. The first is the average number of publications over three years for each scholar (*AvgPublication3*). The second addresses research impact as measured by the number of citation of the papers. Since the number of citations of an article tend to be low for the first few years after publication, we took this time lag effect into consideration and measured the total number of citations over three years. In our data, the citation of an article are collected three years after its publication, at which point the citation score is calculated (i.e. the number of citations in the paper, excluding self-citations). The variable used is the total number of citation scores over three years in our model (*SumCitation3*).

Co-authorship networks and network measures

Our two hypotheses focus on the four classic measures of centrality and cliquishness within co-publication networks mentioned above: degree, closeness and betweenness centralities as well as the clustering coefficient.

[4] The University of Leiden's Centre for Science and Technology Studies. This database was disambiguated by the Leiden scientists according to the algorithm proposed by Caron and van Eck (2014). The resulting database provides a unique author ID for each scientist. As databases are never perfect, we manually checked all records for which we had multiple similar names or multiple email addresses for the same surname – first-name or surname – initials groupings by searching for these individuals on the Internet in order to further disambiguate author names and correct any mistakes in the unique ID, using both email addresses and affiliations. The Centre for Research on Evaluation, Science and Technology (CREST) developed an extensive thesaurus of affiliations, in particular for the African affiliations, that was included in the Leiden database.

N_y denotes a co-authorship network in our model using all co-publication links from 2000-2015. We use intervals within that time period to build the sub-networks, which consist of the co-publication networks anytime from $y_{t-(interval-1)}$ to y_t. Two authors are linked in this network if they have published an article together sometime in this period. In a five-year co-authorship network, which is the case in our study, the interval equals five and any two authors are connected if $Author_i$ publishes an article with $Author_j$ within the period of y_{t-4} to y_t. The network measures are calculated in the same time window for each individual author. To build the panel data necessary for our model, we calculate the network measures for each year and consider the network measures at the end of interval year (in year *t*). For example, for the y_t =2005 the network measures are computed for the period of 2001-2005. We calculate the network measures of individuals in each specific domain of Health, Clinical Medicine and Biomedical Research since the centralities are not comparable without considering the authors specialization.

The centrality and clustering coefficient measures of each individuals in the co-authorship network are calculated using Social Network Analysis (SNA) which maps and measures the relationships among individuals or groups to process the social structure through the use of graph theory. The degree centrality of $Author_k$ is calculated according to Eq.1:

$$DegreeCent(Author_k) = \sum_{i=1}^{n} f(Author_i, Author_k) \qquad \text{Eq.1}$$

where *n* is the number of authors in the network and $f(Author_i, Author_k)$ is a function that equals 1 if these two authors are connected and 0 otherwise (Freeman, 1978). The variable name used in our models for degree centrality is *DegreeCent*, and is defined for each of the three domains.

To measure closeness centrality, a distance function is used. The closeness centrality of $Author_k$ is given by Eq.2 (and the variable name used in our models for closeness centrality is *CloseCent*):

$$CloseCent(Author_k) = 1/\sum_{i=1}^{n} d(Author_i, Author_k) \qquad \text{Eq.2}$$

where *n* is the number of authors in the network and $d(Author_i, Author_k)$ is a distance function (Freeman, 1978).

Betweenness centrality takes the number of shortest paths between two authors in the network into account. Eq.3 presents the formula to compute betweenness centrality (which is referred to in our models by the variable *BetwCent*):

$$BetwCent(Author_k) = \sum_{i}^{n} \sum_{j}^{n} \frac{g_{ij}(Author_k)}{g_{ij}} \; where \; i \neq j \neq k \qquad \text{Eq.3}$$

where g_{ij} defines the total number of geodesics[5] from $Author_i$ to $Author_j$, and $g_{ij}(Author_k)$ represents the number of shortest paths from $Author_i$ to $Author_j$ that passes through $Author_k$ (Freeman, 1978). If the edge e_{ij} connects $Author_i$ with $Author_k$, the clustering coefficient or cliquishness for $Author_k$ is given by the proportion of links between $Author_k$ and the authors within its neighborhoods divided by the maximum number of links between them. Therefore, if $Author_k$ has A_k neighbors, the number of edges that could possibly exist among authors is calculated by $\frac{A_k\,(A_k-1)}{2}$. The clustering coefficient is defined as seen below in Eq.4, where e_k is the number of links between neighbors of $Author_k$ (Hanneman, and Riddle, 2005). The variable name used in our models for the clustering coefficient is *ClusteringC*.

$$ClusteringC(Author_k) = \frac{2e_k}{A_k(A_k-1)} \qquad \text{Eq.4}$$

Control variables

In addition to the network variables of interest described in the previous section, we control for a number of effects that also influence scientific production and impact.

The number of authors is considered as a significant predictor of research productivity. While Lee and Bozeman (2005) found that the number of collaborators in an article affects the publishing productivity, the extent of this effect is dependent upon the productivity of the collaborating scholars. Collaborating with highly-productive scholars increases the quality of a co-authored article; inversely, collaboration with low-productive scholars tend to decrease the quality of that article (Didegah and Thelwall, 2013; Tahmooresnejad and Beaudry, 2017; Aldieri et al., 2018). However, Beaver (2001) highlighted that there are numerous reasons to accept the idea that collaboration has a positive impact on productivity. In co-authored research, different scholars contribute their skills, experiences, and knowledge for the purpose of providing a synergistic effect for the team.[6]

Collaborations with scholars from different countries tend to be associated with higher productivity (Narin et al. 1991; Sooryamoorthy, 2009). We observe that since 2000, African scholars have had an

[5] A geodesic is a path through the network that connect $Author_i$ and $Author_j$.
[6] We first employed the average number of authors in all articles of a scholar in a specific year (*AvgnbAut)* in our model as a control variable. We rapidly realised the importance of the location of the collaborators and therefore only report the results where the number of authors is separated by region (South Africa, Rest of Africa, and International, i.e. outside of the African continent).

increasing tendency to carry out research that includes international collaborations. The increase of these international collaborations over the time magnifies the role of non-African countries listed in "African" publications. In order to take this measure into account, we tested both the number of co-authors affiliated to international institutions (*nbIntAuthors*), i.e. measured here as outside of the African continent[7,8]. We also investigated the collaboration that South African's have with the rest of Africa by calculating the number of the co-authors whose African affiliation is in countries other than South Africa (*nbROAfrAuthors*). To ensure that we account for the total number of authors, we include the number of authors from South Africa (*nbSAAuthors*).[9]

An additional control measures the diversity of the disciplines associated to the authors' papers. For this purpose, we employ the Herfindahl-Hirschman Index (HHI) of the subject categories associated with the journals in which the papers are published. This common measure of market concentration is often used in the literature as an indicator of variety. In our case, the HHI measures the variety of fields in which an author has published articles (*DiversityDiscipline*). HHI equals 1 if the author publications in a given period have the greatest concentration. The formula for this index presented below (Eq. 5) uses the weights of each discipline (measured by the subject category) for each journal

---

[7] Although we tested both the numbers and the proportions of authors per region (*PropIntAuthors*, *PropROAfrAuthors*, *PropSAAuthors*), only the former will be presented in the paper. The results are more reliable because they do not share a common denominator.

[8] It is also relevant for African scientists to collaborate with countries outside of the continent that speak the same language such as English, French, and Arabic. During the course of our analysis, we did not find robust results in this regard. These results are therefore omitted from the paper.

[9] Employing the Herfindahl-Hirschman Index (HHI) logic, we also calculated the diversity of countries (*DiversityCountry*) and institutes (*DiversityInstitutes*) associated with the author list of each paper. In addition to being strongly correlated with one another, with the number of authors, and those from international regions (see the correlation tables in the appendix), these variables did not add any more insight to the regression analysis compared to splitting the number of authors by region. The formulas to calculate both measures are explained below. If all authors' affiliations are from the same country, the calculated country HHI equals 1 and therefore the diversity of countries equals 0. At the other end of the spectrum, if the paper comprises one author per country, then the calculated country HHI equals 1 divided by the number of countries and the diversity of countries is 1 minus that number. A similar variable was built for the diversity of institutes. If all authors are from the same institute, the calculated institute HHI equals 1 and the institute diversity equals 0, and if there is one author per institute, then the calculated institute HHI equals 1 divided by the number of institutes and the diversity of institutes is 1 minus that number.

$$DiversityCountry_p = 1 - \frac{\sum_{c=1}^{N_{cp}} n_c^2}{\left(\sum_{c=1}^{N_{cp}} n_c\right)^2}$$

where n_c is the number of authors of paper p whose affiliation is in country c, and N_{cp} represents the total number of countries listed in the affiliations of paper p.

$$DiversityInstitute_p = 1 - \frac{\sum_{i=1}^{N_{ip}} n_i^2}{\left(\sum_{i=1}^{N_{ip}} n_i\right)^2}$$

where n_i is the number of authors of paper p whose affiliation is i, and N_{ip} represents the total number of different affiliations i associated with the author list of paper p.

in which the papers are published. The disciplinary diversity is calculated using the HHI based on the publications of each author in each specific year:

$$DiversityDiscipline(Author_k) = 1 - \frac{\sum_{d=1}^{N_{dp}} \left(\sum_{p=1}^{N_p} w_d \right)^2}{N_p^2} \qquad \text{Eq.5}$$

Where N_p is the total number of papers p published by author k in a given year, N_{dp} is the total number of subject categories associated with the journal in which the article p is published, and w_d is the weight of each subject category (discipline) d associated with the journal in which the article p is published[10].

Finally, the last control variable used in our models for examining the research impact is the citation score of the journals in which the articles have been published, otherwise known as normalized journal score. We include the average normalized journal citation score over three years (*AvgNJS3*) in our models. According to Liao and Yen (2012), prolific scholars are often selected to publish in high-impact factor journals. Their articles are thus more likely to receive more citations in high impact journals.

Model

The database is built as an unbalanced panel providing data for the years 2000 to 2015 for each individual scientist. With the exception of the diversity of disciplines (*DiversityDiscipline*), all variables are averaged at the individual level for a given year or a group of years (see Table 1 for details). We normalise both our dependent variables using the natural logarithm to be able to use ordinary least squares regressions for panel data. Because individual-level effects are not appropriately modeled by random-effect models, all our regressions were estimated using fixed-effects models[11].

A notable concern in our study is the clear link between the network measures and the number of authors contributing to the scientific production of a given author. The higher the number of coauthors, the more connections a given author has in the copublication network. In addition, the stability of the research teams is such that even using lags for the network measures in the estimations,

---

[10] The Leiden database attributes a number of subject categories to each journal. If a journal covers three subject categories, then a weight (w_d) of 0,33 is attributed to each subject category. Other authors sometimes use forward and backward citations to categorise the diversity of disciplines on which an article is based or used. We chose to use the disciplines associated with the journal itself as more representative of the field in which the authors feel their work belongs.

[11] Hausman specification tests systematically rejected the random effect models in favour of the fixed effect models.

to avoid them being contemporaneous with the number of authors, is not enough to remove the endogeneity inherent in our models. We therefore tested the models to see whether they suffer from potential endogeneity problems. The tests show the presence of endogeneity for the network variables[12]. As individual scientists choose the number of coauthors with whom they collaborate, these selection effects cause the endogeneity issues inherent in these models. To deal with this potential endogeneity a common solution is to employ instrumental variables techniques. Considering the network measures as endogenous, we successfully managed to reduce the effect of the omitted factors by estimating instrumental variable models. Two variables are used in our estimations to control the endogeneity of the network measures. The first, publishing experience, refers to the number of years that a scientist has been active in their scientific network since his or her first publication in the Web of Science. Scientists with a higher *PublishingAge* are more likely to have a wider network of collaborators and to occupy a more central position in their collaboration network. Because we suspected this effect not to be linear over the years, we also added the square of *PublishingAge*.

Based on the assumption that scientific collaborations influence the productivity of scientists, governments have aimed to encourage collaborations. Some of the policies in place are related to the funding of research networks. For instance, in some of the programs in place, scientists are required to collaborate in order to obtain funding from the government. Receiving research funding may thus have a strong power of attraction for collaborators and be associated with a more central role of a scientist in the network. In other words, more funding may signal a high-quality researcher, who we expect should also occupy a key position within his or her scientific networks. We have access to research grants from The National Research Foundation (NRF), which was established on April 1st, 1999, for scientists in South Africa. Although the network and bibliometric measures were built for the African scientists of the entire continent, we restrict the sample for this analysis to scientists from South Africa, but consider their collaborators from all over the world. Our model accounts for the average amount of funding received by an author over 3 years (*AvgFund3*) as an additional instrument for the network measures.

---

[12] All network variables with the exception of closeness centrality are strongly correlated with one another. Each was therefore introduced in the model separately and tested for endogeneity accordingly.

The summary statistics of the variables included in the study over the period 2000 to 2015 are presented in appendix B.

Table 1_Variable Description

Variable	Description
Dependent variables	
$AvgPub3_t$	Average number of publications of the scientist over the past three years. Transformation: [ln($AvgPub3_t$ + 1)]
$SumCit3_t$	Total number of citations received by the publications of the scientist over the past three years. Transformation: [ln($SumCit3_t$ + 1)]
Network variables (endogenous)	
$DegreeCent_t$	Degree centrality of the scientist within the co-publication network for the three Health and Medical Sciences domains. Eq.1. Transformation: [ln(10^3 x $DegreeCent_t$ + 1)]
$CloseCent_t$	Closeness centrality of the scientist within the co-publication network for the three Health and Medical Sciences domains. Eq.2. Transformation: [ln(10^5 x $CloseCent_t$ + 1)]
$BetwCent_t$	Betweenness centrality of the scientist within the co-publication network for the three Health and Medical Sciences domains. Eq.3. Transformation: [ln(10^3x $BetwCent_t$ + 1)]
$ClusteringC_t$	Clustering coefficient of the scientist within the co-publication network for the three Health and Medical Sciences domains. Eq.4. Transformation: [ln(10 x $ClusteringC_t$ + 1)]
Exogenous variables	
$AvgnbIntAut_t$	Average number of international authors (i.e. from outside of the African continent) for the articles published by the scientist in year *t*. Transformation: [ln($AvgnbIntAut_t$ + 1)]
$AvgnbROAfrAut_t$	Average number of African authors from outside of South Africa for the articles published by the scientist in year *t*. Transformation: [ln($AvgnbROAfrAut_t$ + 1)]
$AvgnbSAAut_t$	Average number of South African authors (including the scientist considered) for the articles published by the scientist in year *t*. Transformation: [ln($AvgnbSAAut_t$ + 1)]
$AvgNJS3_t$	Average normalized journal scores of the articles published by the scientist over the past three years. Transformation: [ln($AvgNJS3_t$ + 1)]
$DiversityDiscipline_t$	Average of 1 minus the Herfindahl-Hirschman Index (HHI) of the subject categories associated with each publication of the scientist in year *t.* Eq.5. Transformation: [ln($DiversityDiscipline_t$ + 1)]
Year dummies	A dummy variable for each year accounts for any time trends in a scholar's performance
Instrumental variables	
$AvgFund3_t$	Average amount of funding received by the scientist over the past three years. Transformation: [ln($AvgFund3_t$+ 1)]
$PublishingAge_t$	Number of years since the first publication of the scientist. Transformation: [ln($PublishingAge_t$ + 1)]

Note: Table A in appendix provides a description of the variables tested during the course of the research for which the results are not reported in the paper, but mentioned in some of the footnotes.

Analyses and Results

This section presents the regression results addressing our main question of interest: whether the network position of scholars in the co-authorship network affects their research impact. We tested for potential endogeneity of the network characteristics. The results systematically show that our network variables are endogenous[13]. To overcome this endogeneity, we employ an instrumental technique estimating two stage least squares (2SLS) multivariate regressions for panel data using fixed-effects models (by using the within regression estimator). Table 2 and Table 3 present the second stage of the regression results. Both underidentification and overidentification tests of the instruments[14] confirm that our instruments are valid and uncorrelated with the error term, and that the excluded instruments are relevant and correlated with the endogenous regressors. These results verify the validity of the instruments used in the estimations.

According to the correlation table in the appendix, degree centrality is highly correlated with betweenness centrality and the clustering coefficient. We therefore estimated our models including one network variable at a time which also enables us to capture the endogeneity of all the network variables independently. The discussion of the results first addresses the regressions estimates for the scientific production (see Table 2), followed by that of its impact (Table 3). All network variables are lagged by two-year (in the scientific production model) or three-years (in the citation model) to account for the time lapse between network characteristics and generating scientific output.

Table 2 shows the results of the estimation in which the dependent variable is the average number of publications of scholars over the past three years. The results are presented separately for each of three studied domains. The implication of these regressions is that degree centrality and betweenness centrality in co-authorship networks have a positive impact on the scientific production of scholars, but the coefficient for closeness centrality is not significant in the Health and Clinical Medicine domains, and weakly significant in Biomedical Research. In contrast, the clustering coefficient shows a very strong negative impact, suggesting that the number of articles decreases with excessive cliquishness around a given scientist. Scholars tend to collaborate with others in order to produce

[13] Endogeneity tests were performed in non-panel 2SLS multivariate regressions accounting for the non-independence of observations for the same individual. Hausman specification tests consistently rejected exogeneity of our models.
[14] The Sargan test for overidentifying restrictions was never rejected. Furthermore, the Anderson's canonical correlations were always significantly different from zero, i.e., the Lagrange multiplier statistic χ^2 test was strongly rejected, hence suggesting no underidentification. These test results show that the models estimated with instrumental variables are identified and valid.

more articles, but a higher clustering coefficient hampers scholars' publishing activity, i.e. individual scientists who collaborate with a greater number of non-connected teams (or clusters) have a higher scientific production. When authors collaborate with scholars other than those located within their clustered network, it is believed that they also have access to additional resources from these other groups. This finding is in line with the results of Tahmooresnejad et al. (2015). In terms of the impact of betweenness centrality, it seems that researchers of our sample who are well positioned to serve as intermediaries, and facilitate the flow of information within the network between other researchers, contribute to a greater number of publications. We found that authors whose publications cite a higher number of references are more likely to publish more[15]. Since the number of references is highly correlated with the disciplinary diversity, we tested two different sets of models. The results including the disciplinary diversity are presented in Table 2 and Table 3. They show that authors who publish in a wide variety of disciplines are also those who are the most prolific.

Turning now to the importance of collaborating widely across the world, our results show that South African authors benefit from having a greater number of coauthors from the rest of Africa. Furthermore, if their co-authors are also from South Africa, they are more likely to publish a greater number of papers[16]. Yet, we cannot find a positive effect from the collaboration with international authors on scientific production in our sample. One possible explanation for this impact is that the influence of coauthors from multiple countries may primarily affect the visibility of research output and not the actual production itself. These findings are generally in line with those prior studies (Hollis, 2001; Lee and Bozeman, 2005; Abbasi et al., 2011) that determined that collaborations improve the research performance. In addition, the results of the effect of the journal citation score highlight that when authors publish in higher impact journals, they are also more likely to have a higher number of publications.

The regression results on the factors associated with a greater normalized citations score, which are presented in Table 3, suggest that higher degree and betweenness centrality are associated with a higher citation score in Biomedical Research and Clinical Medicine[16]. The impact of the clustering coefficient on the number of citations is negative. This suggests that scholars collaborating within

---

[15] The regression results including the number of references as opposed to the diversity of disciplines is available upon request from the authors.
[16] We must highlight here the small sample associated with the Health domain (number of individuals = 353 and number of observations = 1179). This explains in a large part, the non-significance of some of the regression coefficients for this sector.

highly clustered networks receive a lower number of citations. This in turns restricts their work being seen by others outside of the network. In order to benefit from the improved visibility associated with a greater number of citations, authors need to pursue new collaborations outside of their clustered network. The increased access to additional and possibly more diversified resources that accrues from widening ones' research cliques, has a positive impact on the resulting quality of their publications.

In addition to the network effects, we also observe that the effect of the number of authors on the citation score is positive and significant for all three regions. The collaboration of South African authors with one another, as well as colleagues from both the rest of Africa and international scholars greatly influence the citation score of their publications. Researchers should therefore value co-authorships rather than single authorships to reach a higher impact level of their publications. In essence, collaborating with a growing number of scholars increases the likelihood of finding suitable co-authors and, thereby, results in increased research impact. In addition to involving far reaching colleagues in one's research, widening the disciplinary diversity (*DiversityDiscipline*) of the journals in which one publishes also provides an interesting opportunity for a higher citation score.

Having published a greater number of articles (*AvgPub3*) in the past is also associated with a higher citation score. As Beaudry and Larivière (2016) suggested, this higher impact is due to the higher visibility that comes with publishing more papers. Similarly, we expected that the researchers benefit from the increased visibility associated with publishing in high-impact factor journals measured by the average normalized journal score. The results show that publications in these prestigious journals are more likely to receive more citations. This effect is very likely a consequence of collaboration with other scholars and mostly from connecting to prolific researchers.

Regarding centrality measures, two facts remain constant: 1) in order to increase the likelihood of a central position, scholars probably need to collaborate with the prolific authors in their group; and 2) to increase their closeness and betweenness centrality, scholars must connect with prominent scholars from other research groups in order to occupy a bridge position between research groups. If these measures are followed, authors may then positively influence their research group and consequently improve the research impact of the entire group. According to our findings, scholars should be encouraged to expand their networks in order to increase the likelihood of participating in collaborative projects. International collaborations outside the African continent are an excellent way for authors to become acquainted with prolific researchers as we observe that a higher number of international co-authors contributes to the papers being noticed and cited. Moreover, publishing in journals associated with various disciplines positively affects the citation scores as the author may be known as an expert in different fields.

To summarize our results, our research shows a positive impact of the network indicators on the performance measures. Significant positive correlations reveal that scholars who are more central in their co-authorship network contribute to more publications. This finding allows us to accept our first hypothesis for three of the four sub-hypotheses (H1a, H1b and H1d)[17]. Authors that collaborate more or who are frequently connecting the paths between other authors have more publications on average.

Furthermore, the correlation between network measures and citation scores shows that strong relationships between co-authors generally result in higher citations. Access to various concepts in a collaboration network provides an opportunity to publish high-quality research. These findings substantiate the second hypothesis and their sub-hypotheses (H2), but only in Biomedical Research and Clinical Medicine. We believe that the low number of observations in Health is the reason that we could not capture the impact of network measures in this domain. Our observations support the findings of a number of scholars (Van Raan, 2004, Persson et al., 2004; Glänzel and Schubert, 2004) who have generally showed that collaboration is associated with an increase in citation counts.

---

[17] The weakly significant positive impact of the closeness centrality in the Biomedical Research domain is not enough to validate H1c.

Table 2– Impact of network measures on research in South Africa in the domains of Health, Biomedical Research and Clinical Medicine, Second stage of 2SLS fixed effect model for panel data

Variables	**Health**				**Biomedical Research**				**Clinical Medicine**			
ln(AvgPub3)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
$ln(10^3 \times DegreeCent_{t-2})$	0.3665 ***				0.5904 ***				0.3262***			
	(0.1162)				(0.0971)				(0.0360)			
$ln(10^5 \times CloseCent_{t-2})$		-0.0999				5.1164 *				0.2919		
		(1.1370)				(2.6744)				(1.8636)		
$ln(10 \times ClusteringC_{t-2})$			-0.4141***				-0.6524 ***				-0.3776 ***	
			(0.1325)				(0.1113)				(0.0449)	
$ln(10^3 \times BetwCent_{t-2})$				0.0369 ***				0.0675 ***				0.0322 ***
				(0.0128)				(0.0120)				(0.0040)
$ln(10^2 \times AvgnbIntAut_t)$	0.0141	0.0143	0.012	0.0114	0.0099 *	-0.0083	0.0064	0.0094 *	0.0031	0.0032	0.0032	0.0035
	(0.0089)	(0.0093)	(0.0087)	(0.0089)	(0.0051)	(0.0120)	(0.0047)	(0.0054)	(0.0028)	(0.0031)	(0.0029)	(0.0029)
$ln(10^2 \times AvgnbROAfrAut_t)$	0.0222 **	0.0224 **	0.0236**	0.0232 **	0.0267 ***	0.0226 *	0.0204 ***	0.0250 ***	0.0279***	0.0313***	0.0288 ***	0.0302 ***
	(0.0095)	(0.0089)	(0.0093)	(0.0095)	(0.0067)	(0.0131)	(0.0063)	(0.0070)	(0.0040)	(0.0041)	(0.0041)	(0.0042)
$ln(10^2 \times AvgnbSAAut_t)$	0.0551	0.0507	0.0349	0.0476	0.0938 ***	0.0154	0.0885 ***	0.0928 ***	0.0799***	0.0775***	0.0660 ***	0.0769 ***
	(0.0412)	(0.0395)	(0.0406)	(0.0411)	(0.0234)	(0.0574)	(0.0217)	(0.0245)	(0.0130)	(0.0153)	(0.0132)	(0.0134)
$DiversityDiscipline_t$	0.3492 ***	0.3448 ***	0.3837***	0.3696 ***	0.3885 ***	0.2818 ***	0.3812 ***	0.3497 ***	0.4201***	0.4325***	0.4422 ***	0.4248 ***
	(0.0667)	(0.0762)	(0.0666)	(0.0673)	(0.0390)	(0.0870)	(0.0363)	(0.0408)	(0.0221)	(0.0253)	(0.0224)	(0.0228)
$ln(AvgNJS3_t)$	0.6358 ***	0.5805 ***	0.5888***	0.5667 ***	0.8237 **	0.9118 ***	0.9177 ***	0.8533 **	0.8013***	0.9021***	0.8503 ***	0.8365 ***
	(0.0728)	(0.1250)	(0.0701)	(0.0716)	(0.0556)	(0.1033)	(0.0466)	(0.0563)	(0.0283)	(0.0346)	(0.0268)	(0.0280)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb of observations	1179	1179	1179	1179	4171	4171	4171	4171	8783	8783	8783	8783
Nb of groups	353	353	353	353	1142	1142	1142	1142	2071	2071	2071	2071
F	13.435 ***	14.698 ***	13.700***	13.382 ***	44.599 ***	11.300 ***	50.872 ***	40.290 ***	169.598***	158.159***	165.385 ***	158.524 ***
R^2	0.1265	0.2356	0.1598	0.1291	-0.0488	-3.0152	0.091	-0.1579	0.3117	0.2829	0.2974	0.2671
Loglikelihood	-454.16	-375.54	-426.82	-452.44	-2437.9	-5237.6	-2137.8	-2644.4	-4056.2	-4236.25	-4128.15	-4331.73

Note: ***, **, * show significance at the 1%, 5% and 10% levels and standard errors are presented in parentheses.

Table 3– Impact of network measures on citations in South Africa in the domains of Health, Biomedical Research and Clinical Medicine, Second stage of 2SLS fixed effect model for panel data

Variables	**Health**				**Biomedical Research**				**Clinical Medicine**			
ln(SumCit3)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
$ln(10^3 \times DegreeCent_{t-3})$	0.2483				0.2297 **				0.1193***			
	(0.1681)				(0.1032)				(0.0419)			
$ln(10^5 \times CloseCent_{t-3})$		-3.2584				0.5426				3.9535		
		(3.1619)				(1.8520)				(3.2904)		
$ln(10 \times ClusteringC_{t-3})$			-0.2571				-0.205 **				-0.123 **	
			(0.1770)				(0.1037)				(0.0480)	
$ln(10^3 \times BetwCent_{t-3})$				0.028				0.0198 *				0.0101 **
				(0.0192)				(0.0104)				(0.0040)
$ln(10^2 \times AvgnbIntAut_t)$	0.0430 ***	0.0354 **	0.0409***	0.0395 ***	0.0372 ***	0.0353 ***	0.0345 ***	0.0359 ***	0.0353***	0.0278***	0.0351 ***	0.0354 ***
	(0.0123)	(0.0159)	(0.0121)	(0.0121)	(0.0063)	(0.0072)	(0.0061)	(0.0062)	(0.0040)	(0.0093)	(0.0040)	(0.0040)
$ln(10^2 \times AvgnbROAfrAut_t)$	0.0548 ***	0.067 ***	0.0561***	0.0515 ***	0.0301 ***	0.0306 ***	0.0293 ***	0.0304 ***	0.025***	0.0248***	0.0246 ***	0.0251 ***
	(0.0131)	(0.0182)	(0.0130)	(0.0139)	(0.0081)	(0.0081)	(0.0080)	(0.0080)	(0.0057)	(0.0085)	(0.0057)	(0.0057)
$ln(10^2 \times AvgnbSAAut_t)$	0.1018 *	0.0829	0.0908	0.1046 *	0.1183 ***	0.1218 ***	0.1162 ***	0.1206 ***	0.1022***	0.0894***	0.0971 ***	0.0988 ***
	(0.0577)	(0.0713)	(0.0564)	(0.0585)	(0.0286)	(0.0280)	(0.0282)	(0.0283)	(0.0186)	(0.0297)	(0.0187)	(0.0186)
$DiversityDiscipline_t$	0.2917 ***	0.3064 **	0.2958***	0.2780 ***	0.3142 ***	0.3091 ***	0.314 ***	0.3004 ***	0.3736***	0.3818***	0.3762 ***	0.3715 ***
	(0.0911)	(0.1191)	(0.0916)	(0.0909)	(0.0468)	(0.0502)	(0.0459)	(0.0460)	(0.0315)	(0.0483)	(0.0315)	(0.0315)
$ln(AvgPub3_{t-1})$	0.1778 ***	0.1637 ***	0.1745***	0.1742 ***	0.1303 ***	0.1335 ***	0.1325 ***	0.1287 ***	0.1268***	0.1238***	0.1294 ***	0.1285 ***
	(0.0177)	(0.0264)	(0.0178)	(0.0179)	(0.0081)	(0.0080)	(0.0078)	(0.0082)	(0.0054)	(0.0096)	(0.0052)	(0.0053)
$ln(AvgNJS3_t)$	1.5961 ***	1.273 ***	1.5664***	1.5648 ***	1.8839 ***	1.8706 ***	1.8789 ***	1.887 ***	1.9031***	1.8802***	1.8909 ***	1.8921 ***
	(0.1094)	(0.2791)	(0.1029)	(0.1028)	(0.0613)	(0.0594)	(0.0600)	(0.0610)	(0.0398)	(0.0582)	(0.0390)	(0.0391)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb of observations	1179	1179	1179	1179	4171	4171	4171	4171	8783	8783	8783	8783
Nb of groups	353	353	353	353	1142	1142	1142	1142	2071	2071	2071	2071
F	56.19 ***	34.54 ***	56.03***	55.84 ***	188.99 ***	199.48 ***	196.63 ***	193.1 ***	388.25***	171.60***	390.07 ***	388.914 ***
R^2	0.5191	0.2179	0.5177	0.5161	0.4921	0.5195	0.512	0.5034	0.4932	-0.1457	0.4957	0.4942
Loglikelihood	-898.8	-1199.5	-900.5	-902.65	-3378.7	-3260.9	-3293.8	-3331.2	-7062.6	-10639.1	-7041.0	-7054.0

Note: ***, **, * show significance at the 1%, 5% and 10% levels and standard errors are presented in parentheses.

Conclusions and discussion

The aim of this paper is to analyze the influence of co-authorship networks on scientific production and its impact in Africa. The world needs the contribution of African scholars affiliated to African research institutions. Quality research in the Health and Medical Sciences in Africa plays a vital role for the world since the share of the global burden of disease in Africa is approximately 20%, but their scientific contributions is far less. It is therefore important to advance the analysis of the factors that improve the productivity and quality of scientific research in Africa. The deep understanding of African scholars about the reality of what happens on the ground brings essential knowledge to international research teams interested in the health science fields specific or not to the particularities of the African continent.

In this paper, we investigated the importance of the research group and deepened the understanding of the benefits that emerge from the collaboration relationships of South African scholars. The approach proposed allows one to understand the effects that collaborations have on authors' scientific production and the citation impact of their research. We characterized collaboration networks by applying social network analysis measures. Our research fills one of the research gaps regarding the lack of studies using detailed network measures of co-authorship for African scientists.

In particular, in order to understand the impact of research collaborations on research, we have investigated two main hypotheses to assess how the position of scientists in a co-authorship network can affect their research output and its impact. We utilized a comprehensive database of articles and authors in Africa in the Health and Medical Sciences between 2000-2015 to construct the network and compute the network measures used in this study. Our results indicate a strong influence of network measures (centrality of cliquishness), as indicators of the extent of their collaboration, on scientific productivity and impact. We found that if a scholar has more connections or higher-quality connections, the likelihood of having more publications or receiving more citations is increased. Occupying central positions in the network enables authors to extend their relationships with co-authors as well as with the co-authors of co-authors, providing opportunities to exchange knowledge that would normally lead to higher performance and higher efficiency of research.

For instance, betweenness centrality illustrate how globally central a scholar is and what level of strategic importance he or she can offer. We identified that it is advantageous for researchers to have a high degree centrality if they wish to benefit from direct, strong connections, and be able to exchange knowledge that will lead to a higher scientific performance and efficiency while also contributing to increasing the visibility of their research. Moreover, our research clearly showed that scholars who maintain relationships outside of a closely clustered research group perform better than authors who collaborate with scholars within the same cluster.

Furthermore, while working with scientists from a more diverse set of countries does not necessarily imply publishing more, it certainly has a strong positive impact on the quality of the publications that result from these international collaborations. The number of internationally co-authored publications have increased at the global level. It appears that the instauration of research support programs between foreign countries and African countries may have paid off; our research measures the effects of these types of international connections on publications. Our results suggest that intra-continent and international collaborations of South African scholars play an important role on the visibility of their work.

While there are only few studies on research collaboration and higher impact from the network analysis viewpoint (such as Abbasi et al., 2011; Catalá-López et al. 2014; Ductor et al., 2014; Tahmooresnejad and Beaudry, 2018; Troshani and Doolin, 2007), there have been several studies using the number of coauthors as an indicator of research collaborations and research output (such as Chung et al., 2009; Glänzel and Schubert, 2004; Katz and Martin, 1997; Lee and Bozeman, 2005). Our findings on South Africa are generally in line with the findings of both groups of studies that examine the correlation between collaborations and research impact using different approaches.

Our results raise an important question for policy implications: while co-authorship collaborations lead to higher performance, individuals still choose their level of co-authorship although it may not be optimal. This is due to the fact that individuals must address a number of constraints regarding collaboration. They must for instance manage the costs associated with developing and maintaining such research collaboration. There is a need for policies to consider uncertainty and organizational constraints that are destined to inducing scholars to participate in co-authored projects.

Furthermore, our findings may shed light on the indicators that have been previously used to measure the research performance since the benefits from collaborative research are likely to be greater. The position of a scholar in the co-authorship network could be used as an indicator of their research impact along with other current indicators. However, further studies are required to examine the various types of collaborations (e.g., when juniors are involved, with whom they choose to publish depending on the quality of their co-authors) (Ductor, 2015).

Finally, we must highlight the limitations to our study. Although we introduced lags within our regression analysis, our results are not entirely free from some adverse causality effects. Another limitation of this study is due to the selection of the sample. We limited the sample of papers to those only including an African affiliation. We are therefore deliberately omitted the relationships that exist outside of the African continent. As a consequence, the centrality and cliquishness measures may overestimate the importance of the African scholars in the respective networks as the regression analysis is only performed on South African scientists. We do not perceive this overestimation to be a major problem however. Somewhat related to this, because we included network measures for scientists within the entire network, some network measures may have been underestimated because the indirect links between research groups provided by international scholars is missing. In addition, the results may have been different if had performed the analysis only for the researchers who are within the giant component of the network. We are nevertheless confident that the results presented are of value to the African research community.

Acknowledgements

The authors acknowledge funding from the International Development and Research Center (IDRC) grant number 107987-001 as well as the Robert Bosch Stiftung grant number 11.5F081.0006.0. We are also grateful to Rodrigo Costas (University of Leiden's Centre for Science and Technology Studies) who provided the data from the Leiden database of the Web of Science. We are also indebted to Johann Mouton, Nelius Boshoff, Robert Tijssen, Jaco Blanckenberg and Charl Swart (Centre for Research on Evaluation, Science and Technology – CREST –, DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy – SciSTIP), Luc Mougeot and Ann Weston (IDRC), Bassirou Diagne (Polytechnique Montréal) for numerous discussions and comments on bibliometric data in Africa and on collaboration within the continent, and to

Carl St-Pierre for his advice on the development of the questionnaire and for performing the comparative testing.

References

Abbasi, A., J. Altmann, and L. Hossain. (2011). Identifying the Effects of Co-authorship Networks on the Performance of Scholars: A Correlation and Regression Analysis of Performance Measures and Social Network Analysis Measures. *Journal of Informetrics* 5 (4): 594–607.

Aldieri, L., Kotsemir, M., & Vinci, C. P. (2018). The impact of research collaboration on academic performance: An empirical analysis for some European countries. *Socio-Economic Planning Sciences*, 62, 13-30

Beaudry, C., & Larivière, V. (2016). Which gender gap? Factors affecting researchers' scientific impact in science and medicine. *Research Policy*, 45(9), 1790-1817.

Beaver, D. (2001). Reflections on Scientific Collaboration (and its Study): Past, Present, and Future – Feature Report, *Scientometrics* 52(3): 365–77

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1), 55-71.

Caron, E., & Van Eck, N.J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In Proceedings of the 19th International Conference on Science and Technology Indicators (pp. 79-86).

Catalá-López, F., Alonso-Arroyo, A., Hutton, B., Aleixandre-Benavent, R., & Moher, D. (2014). Global collaborative networks on meta-analyses of randomized trials published in high impact factor medical journals: a social network analysis. *BMC medicine*, 12(1), 15

Chung, K. H., Cox, R. A., & Kim, K. A. (2009). On the relation between intellectual collaboration and intellectual output: Evidence from the finance academe. *The Quarterly Review of Economics and Finance*, 49(3), 893-916

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, 7(4), 861-873.

Ductor, L., Fafchamps, M., Goyal, S., & van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, 96(5), 936-948.

Ductor, L. (2015). Does co-authorship lead to higher academic productivity?. *Oxford Bulletin of Economics and Statistics*, 77(3), 385-407.

Ductor, L., Goyal, S., & Prummer, A. (2018). Gender & collaboration (No. 856). Working Paper

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3), 215-239.

Glänzel, W. (2001). National characteristics in international scientific co-authorship. Scientometrics, 51, 69–115.

Glänzel W., Schubert A. (2004) Analysing Scientific Networks Through Co-Authorship. In: Moed H.F., Glänzel W., Schmoch U. (eds) *Handbook of Quantitative Science and Technology Research. Springe*r, Dordrecht

Hamermesh, D. S., & Oster, S. M. (2002). Tools or toys? The impact of high technology on scholarly productivity. *Economic Inquiry*, 40(4), 539-555

Hanneman, R., and M. Riddle. 2005. Introduction to Social Network, Methods. Riverside: University of California.

Hollis, A. (2001). Co-authorship and the Output of Academic Economists. Labour Economics 8 (4): 503–530.

Katz, J.S., Martin, B.R. (1997). What is research collaboration? *Research Policy*, 26, 1–18.

Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, 11, 3–15.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social studies of science*, 35(5), 673-702.

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9), 1515-1530.

Liao, C. H., & Yen, H. R. (2012). Quantifying the degree of research collaboration: A comparative study of collaborative measures. *Journal of Informetrics*, 6(1), 27-33

Lindenlaub, I., & Prummer, A. (2020). Network structure and performance. *The Economic Journal.*

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific cooperation in Europe and the citation of multinationally authored papers. *Scientometrics*, 21(3), 313–323.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA, 98, 404–409.

Newman, M. E. (2003). Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences, 101(suppl 1), 5200-5205.

Newman, M.E.J. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. To appear in E. Ben–Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks. Berlin: Springer.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6), 441-453.

Persson, O., Glänzel, W., Danell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies, *Scientometrics*, 60, forthcoming.

Pouris, A., & Ho, Y. S. (2014). Research emphasis and collaboration in Africa. *Scientometrics*, 98(3), 2169-2184

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics*, 81(1), 177–193.

Tahmooresnejad, L. , Beaudry, C. (2018): The importance of collaborative networks in Canadian scientific research, *Industry and Innovation*, DOI: 10.1080/13662716.2017.1421913

Tahmooresnejad, L., & Beaudry, C. Collaboration or funding: lessons from a study of nanotechnology patenting in Canada and the United States (2017). *The Journal of Technology Transfer*, 1-37.

Tahmooresnejad, L., Beaudry, C. & Schiffauerova, A. (2015). The role of public funding in nanotechnology scientific production: Where Canada stands in comparison to the United States, *Scientometrics*, 102: 753.

Troshani, T., Doolin, B. (2007). Innovation diffusion: a stakeholder and social network view. *European Journal of Innovation Management* 10 (2), 176–200.

Van Raan, A. F. (2004). Measuring science. In *Handbook of quantitative science and technology research* (pp. 19-50). Springer, Dordrecht.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107-2118.

Table A – Description of the variables that were tested but for which the results are not presented in the paper

Variable	Description
Exogenous variables	
AvgnbAut$_t$	Average number of authors for the articles published by the scientist in year *t*. Transformation: [ln(*AvgnbAut$_t$*+ 1)]
PropIntAuthors$_t$	Average proportion of international co-authors of the articles of the scientist in year *t*. Transformation: [ln(*PropIntAuthors$_t$* + 1)]
PropROAfrAuthors$_t$	Average proportion of authors from the rest of Africa (other than South African authors) of the articles of the scientist in year *t*. Transformation: [ln(*PropROAfrAuthors$_t$* + 1)]
DiversityCountry$_t$	Average of 1 minus the Herfindahl-Hirschman Index (HHI) of the number of countries associated with the author list of the publications of the scientist in year *t*. Transformation: [ln(*DiversityCountry$_t$* + 1)]
DiversityInstitute$_t$	Average of 1 minus the Herfindahl-Hirschman Index (HHI) of the number of different institutes associated with the author list of the publications of the scientist in year *t*. Transformation: [ln(*DiversityInstitute$_t$* + 1)]

Table B – Descriptive Statistics

Health	Obs	Mean	Std. Dev.	Min	Max
AvgPub3$_t$	1179	*3.2712*	*3.9375*	*0.3333*	*33.3333*
SumCit3$_t$	1179	*222.5168*	*500*	*0*	*5949*
DegreeCent$_t$	1179	*20*	*24*	*2*	*306*
ClusteringC$_t$	1179	*0.4643*	*0.3521*	*0*	*1.0000*
CloseCent$_t$	1179	*0.0002*	*0.0001*	*0.0000457*	*0.0005*
BetwCent$_t$	1179	*86030*	*270946*	*0*	*5209641*
*AvgnbAut$_t$**	1179	*8.0355*	*12.3475*	*1*	*230*
AvgnbIntAut$_t$	1179	*3.4662*	*11.5571*	*0*	*217.7500*
AvgnbROAfrAut$_t$	1179	*0.4078*	*1.2148*	*0*	*22.0000*
AvgnbSAAut$_t$	1179	*4.1615*	*2.1243*	*1*	*15.0000*
AvgNJS3$_t$	1179	*1.0102*	*0.8009*	*0.0237333*	*7.3179*
DiversityDiscipline$_t$	1179	*0.5304*	*0.2896*	*0*	*0.9101*
AvgFund3$_t$	1179	*110355*	*531560*	*0*	*10000000*
PublishingAge$_t$	1179	*1.2172*	*1.0829*	*0*	*2.8332*
*DiversityInstitute$_t$**	1179	*0.5313*	*0.2252*	*0*	*0.9869*
*DiversityCountry$_t$**	1179	*0.3189*	*0.2138*	*0*	*0.9517*
*PropIntAuthors$_t$**	1179	*0.2814*	*0.2276*	*0*	*0.9615*
*PropROAfrAuthors$_t$**	1179	*0.0342*	*0.0895*	*0*	*0.8571*

Note: * represent variables that are mentioned but not used in the paper (see Table A above).

Biomedical Research	Obs	Mean	Std. Dev.	Min	Max
$AvgPub3_t$	4171	*2.4892*	*2.9686*	*0.3333*	*33.3333*
$SumCit3_t$	4171	*151*	*347*	*0*	*7028*
$DegreeCent_t$	4171	*28*	*42*	*2*	*630*
$ClusteringC_t$	4171	*0.3979*	*0.3197*	*0.0000*	*1.0000*
$CloseCent_t$	4171	*0.0001*	*0.0000*	*0.0000*	*0.0001*
$BetwCent_t$	4171	*518681*	*1989995*	*0*	*66600000*
$AvgnbAut_t^*$	4171	*8*	*17*	*1*	*890*
$AvgnbIntAut_t$	4171	*3*	*16*	*0*	*867*
$AvgnbROAfrAut_t$	4171	*0*	*1*	*0*	*26*
$AvgnbSAAut_t$	4171	*4*	*2*	*1*	*23*
$AvgNJS3_t$	4171	*0.8578*	*0.7245*	*0.0071*	*11.3467*
$DiversityDiscipline_t$	4171	*0.4986*	*0.3000*	*-3.0000*	*0.9359*
$AvgFund3_t$	4171	*125619*	*473269*	*0*	*10200000*
$PublishingAge_t$	4171	*0.9918*	*1.0700*	*0.0000*	*2.8332*
$DiversityInstitute_t^*$	4171	*0.4563*	*0.2481*	*0.0000*	*0.9922*
$DiversityCountry_t^*$	4171	*0.2737*	*0.2229*	*0.0000*	*0.9307*
$PropIntAuthors_t^*$	4171	*0.2450*	*0.2396*	*0.0000*	*0.9805*
$PropROAfrAuthors_t^*$	4171	*0.0228*	*0.0805*	*0.0000*	*0.8889*

Note: * represent variables that are mentioned but not used in the paper (see Table A above).

Clinical Medicine	Obs	Mean	Std. Dev.	Min	Max
$AvgPub3_t$	*8783*	*2.2218*	*2.5114*	*0.3333*	*33.3333*
$SumCit3_t$	*8783*	*141*	*338*	*0*	*7028*
$DegreeCent_t$	*8783*	*47*	*68*	*2*	*1029*
$ClusteringC_t$	*8783*	*0*	*0*	*0*	*1*
$CloseCent_t$	*8783*	*0.000046*	*0.000010*	*0.000007*	*0.000064*
$BetwCent_t$	*8783*	*1614318*	*5277479*	*0*	*123000000*
$AvgnbAut_t^*$	*8783*	*8*	*20*	*1*	*890*
$AvgnbIntAut_t$	*8783*	*3*	*19*	*0*	*867*
$AvgnbROAfrAut_t$	*8783*	*0*	*1*	*0*	*32*
$AvgnbSAAut_t$	*8783*	*4*	*3*	*1*	*25*
$AvgNJS3_t$	*8783*	*0.8469*	*0.7351*	*0.0048*	*11.3467*
$DiversityDiscipline_t$	*8783*	*0.4468*	*0.3058*	*0.0000*	*0.9359*
$AvgFund3_t$	*8783*	*86503*	*369675*	*0*	*10200000*
$PublishingAge_t$	*8783*	*1.0122*	*1.0532*	*0.0000*	*2.8332*
$DiversityInstitute_t^*$	*8783*	*0.4668*	*0.2498*	*0.0000*	*0.9941*
$DiversityCountry_t^*$	*8783*	*0.2749*	*0.2287*	*0.0000*	*0.9707*
$PropIntAuthors_t^*$	*8783*	*0.2433*	*0.2415*	*0.0000*	*0.9954*
$PropROAfrAuthors_t^*$	*8783*	*0.0221*	*0.0806*	*0.0000*	*0.8889*

Note: * represent variables that are mentioned but not used in the paper (see Table A above).

Appendix B1 – Correlation Matrices

Health		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$AvgPub3_t$	1	1.00																	
$SumCit3_t$	2	0.80	1.00																
$DegreeCent_t$	3	0.38	0.29	1.00															
$ClusteringC_t$	4	-0.22	-0.10	-0.70	1.00														
$CloseCent_t$	5	0.11	0.38	0.20	-0.08	1.00													
$BetwCent_t$	6	0.32	0.19	0.71	-0.71	0.19	1.00												
$AvgnbAut_t^*$	7	0.31	0.39	0.28	0.09	0.08	0.03	1.00											
$AvgnbIntAut_t$	8	0.35	0.42	0.28	-0.03	0.15	0.12	0.63	1.00										
$AvgnbROAfrAut_t$	9	0.32	0.25	0.29	-0.11	0.04	0.18	0.47	0.31	1.00									
$AvgnbSAAut_t$	10	0.20	0.19	0.09	0.13	0.02	-0.05	0.45	-0.01	-0.02	1.00								
$AvgNJS3_t$	11	0.51	0.66	0.22	-0.06	0.15	0.14	0.43	0.37	0.25	0.16	1.00							
$DiversityDiscipline_t$	12	0.51	0.38	0.24	-0.11	0.00	0.17	0.20	0.27	0.19	0.18	0.19	1.00						
$DiversityInstitute_t^*$	13	0.11	0.23	0.23	0.01	0.17	0.05	0.57	0.63	0.32	0.09	0.27	0.08	1.00					
$DiversityCountry_t^*$	14	0.13	0.24	0.21	-0.04	0.17	0.07	0.48	0.80	0.42	-0.23	0.29	0.08	0.69	1.00				
$AvgFund3_t$	15	0.14	0.09	0.05	-0.07	-0.03	0.04	-0.03	-0.04	0.05	0.02	0.02	0.07	-0.08	-0.08	1.00			
$PublishingAge_t$	16	-0.23	-0.29	-0.01	-0.05	-0.24	0.01	-0.02	-0.08	-0.02	0.03	-0.14	-0.06	-0.05	-0.06	-0.09	1.00		
$PropIntAuthors_t^*$	17	0.15	0.27	0.21	-0.04	0.12	0.08	0.46	0.84	0.20	-0.35	0.29	0.10	0.60	0.84	-0.07	-0.07	1.00	
$PropROAfrAuthors_t^*$	18	0.06	0.00	0.13	-0.07	0.01	0.07	0.19	0.08	0.76	-0.18	0.04	0.03	0.23	0.35	-0.01	0.05	0.01	1.00

Note: * represent variables that are mentioned but not used in the paper (see Table A above).

Biomedical Research		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$AvgPub3_t$	1	1.00																	
$SumCit3_t$	2	0.78	1.00																
$DegreeCent_t$	3	0.54	0.47	1.00															
$ClusteringC_t$	4	-0.43	-0.30	-0.69	1.00														
$CloseCent_t$	5	0.07	0.28	0.21	-0.13	1.00													
$BetwCent_t$	6	0.41	0.33	0.71	-0.67	0.26	1.00												
$AvgnbAut_t^*$	7	0.23	0.32	0.42	0.02	0.07	0.19	1.00											
$AvgnbIntAut_t$	8	0.31	0.41	0.38	-0.05	0.15	0.19	0.63	1.00										
$AvgnbROAfrAut_t$	9	0.26	0.21	0.22	-0.07	0.00	0.17	0.29	0.23	1.00									
$AvgnbSAAut_t$	10	0.13	0.08	0.23	0.00	0.01	0.11	0.47	-0.03	-0.07	1.00								
$AvgNJS3_t$	11	0.52	0.64	0.34	-0.10	0.06	0.22	0.44	0.40	0.18	0.14	1.00							
$DiversityDiscipline_t$	12	0.38	0.29	0.25	-0.15	-0.01	0.23	0.20	0.24	0.16	0.13	0.20	1.00						
$DiversityInstitute_t^*$	13	0.10	0.23	0.22	0.08	0.11	0.09	0.53	0.66	0.28	-0.02	0.28	0.09	1.00					
$DiversityCountry_t^*$	14	0.12	0.27	0.23	0.03	0.15	0.08	0.51	0.84	0.35	-0.23	0.31	0.08	0.73	1.00				
$AvgFund3_t$	15	0.13	0.09	0.12	-0.18	0.01	0.17	-0.07	-0.01	0.00	-0.05	-0.03	0.08	-0.09	-0.06	1.00			
$PublishingAge_t$	16	-0.14	-0.19	-0.04	0.08	-0.16	-0.06	0.12	-0.03	0.04	0.18	-0.01	-0.04	-0.02	-0.02	-0.13	1.00		
$PropIntAuthors_t^*$	17	0.14	0.29	0.25	0.01	0.12	0.09	0.49	0.85	0.14	-0.35	0.33	0.10	0.66	0.88	-0.02	-0.05	1.00	
$PropROAfrAuthors_t^*$	18	0.06	0.02	0.04	0.01	-0.01	0.05	0.10	0.05	0.77	-0.21	0.02	0.04	0.20	0.28	-0.05	0.04	0.01	1.00

Note: * represent variables that are mentioned but not used in the paper (see Table A above).

Clinical Medicine		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$AvgPub3_t$	1	1.00																	
$SumCit3_t$	2	0.74	1.00																
$DegreeCent_t$	3	0.75	0.65	1.00															
$ClusteringC_t$	4	-0.69	-0.50	-0.74	1.00														
$CloseCent_t$	5	0.13	0.27	0.26	-0.18	1.00													
$BetwCent_t$	6	0.43	0.32	0.45	-0.30	0.04	1.00												
$AvgnbAut_t^*$	7	0.17	0.31	0.43	-0.03	0.13	0.17	1.00											
$AvgnbIntAut_t$	8	0.28	0.42	0.40	-0.12	0.13	0.16	0.59	1.00										
$AvgnbROAfrAut_t$	9	0.22	0.17	0.26	-0.09	0.02	0.22	0.31	0.22	1.00									
$AvgnbSAAut_t$	10	0.08	0.05	0.24	-0.05	0.12	0.05	0.47	-0.08	-0.06	1.00								
$AvgNJS3_t$	11	0.50	0.64	0.51	-0.31	0.12	0.20	0.40	0.36	0.19	0.11	1.00							
$DiversityDiscipline_t$	12	0.42	0.35	0.36	-0.24	0.06	0.20	0.20	0.25	0.16	0.14	0.20	1.00						
$DiversityInstitute_t^*$	13	0.03	0.20	0.21	0.06	0.11	0.07	0.52	0.63	0.27	-0.02	0.23	0.08	1.00					
$DiversityCountry_t^*$	14	0.10	0.29	0.24	-0.01	0.10	0.08	0.48	0.85	0.34	-0.26	0.30	0.10	0.71	1.00				
$AvgFund3_t$	15	0.11	0.06	0.05	-0.05	-0.04	0.10	-0.02	0.01	0.00	-0.01	-0.02	0.07	-0.06	-0.05	1.00			
$PublishingAge_t$	16	-0.11	-0.19	-0.06	0.11	-0.09	-0.08	0.10	-0.05	0.03	0.16	0.00	-0.01	0.00	-0.04	-0.05	1.00		
$PropIntAuthors_t^*$	17	0.11	0.31	0.23	-0.01	0.06	0.08	0.45	0.85	0.14	-0.39	0.30	0.11	0.64	0.89	-0.02	-0.07	1.00	
$PropROAfrAuthors_t^*$	18	0.05	0.01	0.08	0.00	0.00	0.06	0.14	0.06	0.79	-0.17	0.06	0.04	0.20	0.28	-0.04	0.04	0.01	1

Note: * represent variables that are mentioned but not used in the paper (see Table A above).