This item is the archived peer-reviewed author-version of:

Detecting the emergence of new scientific collaboration links in Africa : a comparison of expected and realized collaboration intensities

Reference:

Guns Raf, Wang Lili.- Detecting the emergence of new scientific collaboration links in Africa : a comparison of expected and realized collaboration intensities
Journal of informetrics - ISSN 1751-1577 - 11:3(2017), p. 892-903
Full text (Publisher's DOI): <https://doi.org/10.1016/J.JOI.2017.07.004>
To cite this reference: <https://hdl.handle.net/10067/1450680151162165141>

This is a postprint of an article published in *Journal of Informetrics*. Please cite as follows:

Guns, Raf & Wang, Lili (2017). Detecting the emergence of new scientific collaboration links in Africa: A comparison of expected and realized collaboration intensities. *Journal of Informetrics*, 11(3), 892–903. <https://doi.org/10.1016/j.joi.2017.07.004>

Detecting the emergence of new scientific collaboration links in Africa: A comparison of expected and realized collaboration intensities

Raf Guns^a and Lili Wang^b

^a University of Antwerp, Centre for R&D Monitoring (ECOOM), Middelheimlaan 1, 2020 Antwerpen, Belgium

^b UNU-MERIT, Boschstraat 24, 6211 AX, Maastricht, The Netherlands

Abstract

The evolution of a collaboration network is to some extent steered by the network topology itself. This is the reason behind the success of network evolution models and approaches to link prediction. At the same time, some changes are due to exogenous factors (i.e., factors external to the network itself). In this paper, we explore changes in the collaboration network of African countries (2000–2014), with a focus on detecting emergent links beyond endogenous network features. Using link prediction and machine learning, we generate an ‘expected’ (predicted) collaboration network based on past data and compare it with the actual network that evolved in later years. The results show that the intensity of collaboration with non-African countries is higher than expected, especially for countries that are scientifically more active. To a lesser extent, we also find an increase in collaboration within Africa, which seems mostly due to the scientifically less developed countries. Emergent collaborations are mostly found in the first half of the studied period; in the second half the network structure appears to stabilize.

1 Introduction

Lack of research capabilities is a well-known factor impeding the growth of scientific research in Africa. Given that most African countries have difficulties in reaching their R&D targets, i.e. 1% of their respective gross domestic product (AOSTI, 2014), relying on project funding alone may hamper the scientific progress in Africa. International collaborations have become one of the most efficient means for capacity building (AOSTI, 2014). International human resources (i.e. international collaborating scholars) are believed to be of great importance in shaping the research agendas in Africa (Chataway, Smith, & Wield, 2005; Toivanen & Ponomarev, 2011). At the same time, a research agenda that is dictated mainly by international collaboration may have little societal relevance (Kreimer & Meyer, 2008). As pointed out by Arvanitis, Waast and Gaillard (2000), due to

the lack of science policies and the “privatization” of science, scientists in many African countries spend much more time in writing contract proposals and doing contracting work rather than conducting original and fundamental scientific research.

Moed and Halevi (2014) propose a four-stage model of scientific development in which the second stage – building up – is especially characterized by a large share of international collaboration. In the building-up phase of a country’s scientific development, international scientific collaborations are often established with developed countries (Moed, 2016). The international links between Africa and other countries are often mediated through cooperative health and agricultural programs (Adams et al., 2014).

Different collaboration structures indicate different types of knowledge flows. Hence the evolution of a network signals the changes of knowledge resources in the process of collaborating and learning. Although the development of a network is sometimes believed to result from earlier connections (Powell, et al. 2005; Albert & Barabási, 2000), exogenous factors can also direct future collaboration links (Boshoff, 2010; Schubert & Sooryamoorthy, 2010). This raises the question to what extent the evolution of the collaboration network of African countries can be explained by endogenous factors (network topology in the past) and how exogenous factors affect the network at the level of (1) the continent as a whole, (2) geographical groups within Africa, and (3) African countries.

This study aims to explore the changes in the collaboration network of African countries, with a focus on detecting emergent links beyond endogenous network features. Using link prediction and machine learning, we generate an ‘expected’ (predicted) collaboration network based on past data and compare it with the actual network that evolved in later years. Hence we show to what extent the network has developed based on existing links. Where there are clear deviations from our predictions, we examine where the main emerging collaboration links are located.

This study makes the following methodological contributions:

- Combination of predictors through random forest regression for predicting link weights;
- Use of link prediction techniques to generate ‘expected’ networks, as a baseline against which to compare actual networks;
- Definition and use of international collaboration intensity as an indicator at the level of countries and geographical groups of countries (geo-groups for short).

2 Literature

2.1 Network prediction

Social networks grow and evolve according to certain regularities. Mechanisms like assortativity, triadic closure, and cumulative advantage (De Solla Price, 1976) have given rise to the existence of different network (growth) models, such as the preferential attachment model (Barabási & Albert, 1999). This indicates that the development of a network is not entirely random and future interactions can often be explained by current network topology, without further information on the actors or social or geographical context. We will refer to these features as *intrinsic* or *endogenous* features of the network, by which we mean that the network structure itself ‘encourages’ the development of the network in a certain direction (e.g., the establishment of a new link).

Predicted and recommended network connections have been discussed in earlier literature (e.g., Guns & Rousseau, 2014; Liben-Nowell & Kleinberg, 2007; Yan & Guns, 2014; Zhou, Lü, & Zhang, 2009). Liben-Nowell & Kleinberg (2007) propose to use proximity features for the prediction of future links. Guns & Rousseau (2014) present a method for recommending research collaboration partners which significantly improves the accuracy of recommendations. Comparing author-, institution-, and country-level collaboration networks, Yan & Guns (2014) find that collaboration at higher levels of aggregation, especially country-level collaboration networks, tend to yield more accurate prediction outcomes than lower-level ones.

Despite the importance of network topology to the evolution of a network, there are also exogenous factors influencing connections. Provided there are new resources and opportunities, actors are expected to alter their positions and reconfigure the networks (Powell et al., 2005). The propensity of one country's engagement in collaboration can be steered by government mediation, strength of the partners, professional recognition, funding possibilities and infrastructure (Boshoff, 2010; Schubert & Sooryamoorthy, 2010; Wagner & Leydesdorff, 2005).

The existence of such exogenous factors implies that link prediction has inherent limits. Our use of link prediction in this paper is somewhat similar to that of Rattigan and Jensen (2005), who propose exploiting these limits to find 'anomalous' links. However, our contribution moves in several ways beyond that of Rattigan and Jensen (2005): we propose a method to aggregate multiple predictions, take link weight into account (see the methods section for details) and explore exogenous factors at the level of nodes (countries) or groups of nodes (geo-groups or even the whole of Africa).

2.2 Scientific collaboration in Africa

Science production in Africa is characterized by its high intensity of international collaborations and unbalanced partnerships (AOSTI, 2014; Boshoff, 2010; Confraria, Godinho, & Wang, 2017; Jacobs & Pichappan, 2006; Mègnigbèto, 2013; Sooryamoorthy, 2009) in particular with a low collaboration frequency between African countries and a high collaboration frequency with non-African countries.

Collaborative networks are widely recognized to be helpful in providing knowledge spillovers and for African countries to advance their innovation systems (Toivanen & Ponomariov, 2011). Using a set of publications of South African scientists, Sooryamoorthy (2009) reports that collaboration has a positive impact on the productivity of South-African scientists. The collaborative partners from other geographical locations form sources of knowledge for the African countries to learn from. Several scholars have discussed internal and external scientific collaborations in Africa (Adams, King, & Hook, 2010; Toivanen & Ponomariov, 2011). Toivanen & Ponomariov (2011) state that there is "potential for important changes under way" for collaboration in Africa (p. 491). It is still largely unknown, however, how and in which direction such collaboration links evolve over time. A dynamic perspective to capture network changes, in our view, is key to deepen the understanding of scientific production in Africa.

2.3 Formation of collaboration networks in Africa

Schubert & Sooryamoorthy (2010) find that the emergence of collaboration between South Africa and Germany is shaped by a centre-periphery pattern. Relying on the scientific output between 2000 and 2006, Boshoff (2009) illustrates that 35 per cent of Central Africa's research papers have been

produced in collaboration with past colonial rulers. Mênigbêto (2013) finds that the major drivers of collaboration for West Africa are language, culture and colonial ties.

Collaborating with advanced partners can provide the opportunity to learn more advanced knowledge and technology and to share expensive equipment. In exchange, local demand and knowledge in developing countries can make the thematic programmes more meaningful (Kreimer & Meyer, 2008). Earlier literature has shown that South Africa is expected to benefit more from collaborating with core players rather than other African countries (Boshoff, 2010; Schubert & Sooryamoorthy, 2010). However, a certain domestic capability is believed to be crucial in absorbing the knowledge transferred from the more advanced partners (Boshoff, 2010). In this regard, collaborations between African countries within the same region and with similar environmental conditions can have the advantage of technological congruence and less “institutional mismatch” barriers (Georghiou, 1998). Moreover, collaboration with First World partners carries the danger that the research priorities are exclusively set by the non-African partners (Kreimer & Meyer, 2008).

It is well acknowledged that a deeper understanding of changes in collaboration structures is of great importance to identify the dynamics of knowledge exchange (Wagner & Leydesdorff, 2005). Despite the increasing attention to scientific collaborations in Africa, to the best of our knowledge, little work has been done in examining the network formation and evolution. By applying network prediction techniques, this study explores the dynamic features of collaboration networks in Africa. Furthermore, to shed light on the great heterogeneity of African sub-groups and their preference for different collaboration partners, special attention will be paid to comparing the collaboration patterns of different African sub-groups and the geographical locations of the partners involved.

3 Data

All data were collected from Elsevier’s Scopus, including all publication types and all academic fields. It should be mentioned at this point that this choice inevitably has an effect on the results. Ocholla, Mostert, and Rotich (2016) report that the visibility of two South-African universities is greater in Scopus than in Web of Science (WoS). The comprehensive comparison of WoS and Scopus by Mongeon and Paul-Hus (2016) shows that Scopus is subject to the same biases as WoS, in spite of its larger journal coverage. This bias is hard to avoid for international comparative studies but should be kept in mind when interpreting the results. The dataset covers all publications of the 54 African countries between 2000 and 2014. Following the definition of the United Nations¹, we classify groups by geographical locations, namely Africa South, North, Central, East and West. The list of countries as well as their geographical groups is provided in Table A1 in the Appendix. All countries geographically located in Africa are treated as African; e.g., Réunion and Mayotte are overseas departments of France but considered to be African in this study. In calculating intensity of collaborations with non-African countries, we focus on the countries that publish jointly with Africa more than 10 articles. Thus there are 50 non-African collaborative countries in 2000, 79 in 2007, and 105 in 2014.

Data from 2000, 2007, and 2014 were transformed into weighted collaboration networks. Each node in the network represents a country (either African or non-African) and each link represents collaboration between the two connected countries, with the link weight denoting the joint number

¹ <http://unstats.un.org/unsd/methods/m49/m49regin.htm#ftna>

of publications. No fractionalization is applied. Table 1 provides some basic descriptive statistics. The growth of the network in terms of nodes, but also density and link weight, is clear. Data from 2005 and 2012 were used to test the prediction robustness.

Table 1. Descriptive statistics of three networks

	2000 network	2007 network	2014 network
# nodes	100	132	159
# links	1756	4119	8033
Median link weight	2	3	5
Density	0.35	0.48	0.64

4 Methods

As shown by Yan & Guns (2014), predictions of collaboration networks at country level produce more accurate results than those at lower levels of aggregation. In this paper, the predictions and analysis are carried out at the level of countries.

As in most link prediction studies, the network topology of an older network – the training network – is used to predict links in a later network – the test network. The 2000 network serves as the training network for the 2007 test network. At the same time, the 2007 network serves as the training network for the 2014 test network.

First, the two networks are pre-processed: countries that do not occur in both the training and the test network or that do not collaborate with any other country in either training or test network were removed. The reason for this is simply that it is not possible to make any predictions concerning countries that are not in the training network. For 2000 and 2007 there are 33 cases, four of which are African countries (Djibouti, Equatorial Guinea, Mayotte, and Somalia). For 2007 and 2014 there are 28 cases, two of which are African countries (Liberia and Somalia).

Subsequently, for each pair of nodes in the training network, we calculate a number of predictors. A predictor is an indicator of the relatedness of two nodes; all of the predictors used have been shown in previous research to have some predictive power to predict future links.

4.1 Predictors

This section provides an overview of the predictors used and their parameters. Four predictors were previously used by Guns and Rousseau (2014). We briefly discuss each predictor in turn. Table 2 provides an overview.

First, let us introduce some notations. The set of neighbours of node u is denoted by N_u . We use $|\cdot|$ to denote set cardinality and, hence, $|N_u|$ is the degree of node u , i.e. the number of nodes it is connected to. The weight of the link between u and v is written as $w(u, v)$.

Common neighbours weighted is defined as follows:

$$p(u, v) = \sum_{z \in N_u \cap N_v} w(u, z) \cdot w(z, v) \quad (1)$$

This takes into account the number of common neighbours of nodes u and v , as well as the strength of the respective links. A slightly different weighted variation of common neighbours was used by Murata and Moriyasu (2007).

The cosine measure is defined as

$$p(u, v) = \frac{|N_u \cap N_v|}{\sqrt{|N_u|} \sqrt{|N_v|}} \quad (2)$$

Resource allocation was introduced by Zhou et al. (2009) and is similar to the more frequently used Adamic/Adar measure. It is a refinement of common neighbours that attaches more importance to ‘rare’ (i.e., low-degree) neighbouring nodes.

$$p(u, v) = \sum_{z \in N_u \cap N_v} \frac{1}{|N_z|} \quad (3)$$

SimRank is a measure of how similar two nodes in a network are, originally proposed by Jeh and Widom (2002) and further elaborated by Antonellis, Molina, and Chang (2008). The SimRank thesis can be summarized as: *Objects that link to similar objects are similar themselves*. We use the weighted SimRank (Antonellis et al., 2008; Guns & Rousseau, 2014):

$$sim(u, v) = \frac{c}{\sum_{p \in N_u} w(u, p) \cdot \sum_{q \in N_v} w(v, q)} \sum_{p \in N_u} \sum_{q \in N_v} sim(p, q) \cdot w(u, p) \cdot w(v, q) \quad (4)$$

where c ($0 < c < 1$) is the ‘decay factor’ that determines how quickly similarities decrease. If, for example, two countries x and y both collaborate with z , then c determines the certainty with which we can state that x and y are similar.

Rooted PageRank is a form of personalized PageRank (Langville & Meyer, 2005). The intuition behind rooted PageRank (Liben-Nowell & Kleinberg, 2007) can be explained using the concept of a random walker that starts at a fixed node v , called the root node. At each step, the walker either moves along a link to a (randomly chosen) neighbour of the current node or is teleported back to the root node v . Teleportation occurs with probability $1 - \alpha$ (where $0 < \alpha < 1$). We use a weighted form of rooted PageRank, such that a link with a higher weight will be more likely to be traversed than a link with a lower weight.

Graph distance weighted is defined as (Guns & Rousseau, 2014):

$$p(u, v) = \left(\sum_{i=1}^t \frac{1}{w_i} \right)^{-1} \quad (5)$$

where w_i ($i = 1, \dots, t$) denotes the weight of the i th link in the shortest path with length t from u to v .

Finally, we turn to Katz weighted, an extension of the Katz measure (Katz, 1953) to weighted networks. Let A denote the adjacency matrix of the network, in which each element a_{ij} is equal to the link weight of nodes v_i and v_j , or 0 if no link is present. Then, $a_{ij}^{(k)}$ is an element from A^k , the k -th power of A . It can be shown that the value of $a_{ij}^{(k)}$ is the number of walks from v_i to v_j with length k in the multigraph described by A (Guns & Rousseau, 2014). The predictor is

$$p(v_i, v_j) = \sum_{k=1}^{\infty} \beta^k a_{ij}^{(k)} \quad (6)$$

where β is a parameter between 0 and 1, that controls the influence of nodes further away.

Table 2. Overview of predictors

Predictor	Parameters
Common neighbours weighted	–
Cosine	–
Resource allocation	–
SimRank weighted	$c = 0.8$
Katz weighted	$\beta = 0.01$
Graph distance weighted	–
Rooted PageRank weighted	$\alpha = 0.5$

4.2 Aggregation by random forest regression

Having a set of different predictors at our disposal immediately poses the problem of which one should be chosen. Since we want to use the prediction as a kind of ‘expected’ value, it is important that we avoid idiosyncrasies of individual predictors. For this reason, the results of the seven predictors have been aggregated into one ‘meta-prediction’ using machine learning.

Since the networks are collaboration networks at the national level, they are quite dense (Table 1). In the 2014 network, 64% of all node pairs are linked. Hence we choose to predict the *weight* of each link (where a weight of 0 implies no link) rather than simply the *occurrence* of a link.

The specific machine learning technique used is random forest regression (Breiman, 2001). A random forest is an ensemble of decision trees, where each tree is based on an independent random sample of the input data. Moreover, each node (i.e., each decision) in a tree is based on a random subset of the available features. The final score is the average of the scores predicted by all trees. By introducing randomness at both the data and the model level, random forests have been shown to yield accurate and robust results. We use random forest regression rather than random forest classification (as was done by Guns and Rousseau, 2014) since there is no theoretical upper bound to link weight.

The random forest regressor was trained on the combined 2000 and 2007 data. To avoid overfitting, 5-fold cross-validation was used, in which 150 trees and a maximum tree depth of 3 were found to be the hyperparameters yielding the best results. This step uses the implementation of Scikit-learn (Pedregosa et al., 2011).

4.3 Collaboration intensity

Given the variety in collaboration patterns for countries in Africa (Adams et al., 2014), this paper takes geographical groups into consideration, i.e. we distinguish between Africa South, North, Central, East and West.

The international collaboration intensity of African country i with partner country j (ICI_{ij}) is defined as the ratio of their number of joint publications (IC_{ij}) to the number of joint publications of country i with all partners in the whole network in the same year. This can be expressed as follows:

$$ICI_{ij} = \frac{IC_{ij}}{\sum_k IC_{ik}} \quad (7)$$

From (7), it follows that, generally speaking, $ICI_{ij} \neq ICI_{ji}$. The sum of international collaboration intensities of a country with all other countries equals 1. We apply formula (7) to both actual and predicted publication counts.

International collaboration intensity (ICI) is further differentiated at three layers:

- Within geo-group collaborations are collaborations with countries from the same African geographical group, that is, i and j belong to the same African geographical group.
- Within Africa collaborations are collaborations with countries in Africa, that is, i and j are both African countries.
- Outside Africa collaborations are collaborations with countries outside of Africa, that is, $i \in$ African countries; $j \in$ non-African countries.

5 Results

5.1 Collaboration predictions

The random forest regressor learns the relative importance of each predictor from the training data. Table 3 displays the results.

Table 3. Relative importance of predictors in final prediction

Predictor	Feature importance
Resource allocation	0.20
Katz weighted	0.18
Rooted PageRank weighted	0.16
Graph distance weighted	0.16
SimRank weighted	0.13
Cosine	0.10
Common neighbours weighted	0.06

Although it is not the main focus of our paper, we briefly present some results to illustrate how our random forest model works on new data. For this, we apply it to the network of 2005 in order to predict link weights in 2012. The resulting R^2 score is 0.60, which illustrates that the model can capture overall trends. Plotting predicted against actual link weights, we see that there is a clear linear correlation between the two but that the prediction tends to underestimate link weights

(Figure 1). A perfect prediction would follow the dotted line. The largest outlier is the number of collaborations between Saudi Arabia and Egypt, which was almost four times higher than predicted.

Figure 1. Comparison of actual and predicted link weights (training network: 2005, test network: 2012)

5.2 Comparison of predicted and actual collaboration intensities

Figure 2 provides a detailed comparison between predicted and actual collaboration intensities. The vertical axis represents the actual collaboration intensity and the horizontal axis represents the predicted collaboration intensity. International collaborations are distinguished at three different layers, within geo-group collaborations in (a) and (b); within Africa collaborations in (c) and (d); and outside Africa collaborations in (e) and (f). For each type of collaboration, the comparison is made at two temporal stages. Namely, the actual 2007 collaboration links are compared with the predicted links based on the 2000 training network; and the actual 2014 collaboration links are compared with the predicted links based on the 2007 training network.

Observations above the (red dotted) diagonal line are the countries that developed a higher collaboration intensity than predicted. Observations below the diagonal line represent the countries whose international collaboration intensity is less than expected.

In Fig. 2, the top two panels provide the comparison for collaborations with countries from the same African geographical group. In panel (a), the observations are located around the diagonal line. This indicates that, on average, the actual international collaborations within African geo-groups did not greatly deviate from the predicted situation. Countries like Lesotho (no. 38), Swaziland (no. 41), Burundi (no. 17) and Comoros (no. 18) are the few exceptions that have greatly advanced their international collaboration intensities within their geo-groups. At a later stage, as shown in panel (b), most countries are located above the diagonal line. This implies that the actual collaboration

intensities between African countries from the same geographical groups in 2014 were generally higher than predicted on the basis of the network of 2007. In other words, most African countries developed within geo-group collaboration links in 2014 beyond what could be expected based on the intrinsic features of the 2007 network. Notably, countries in Western Africa – e.g. Togo (no. 16), Guinea (no. 7) and Guinea-Bissau (no. 8) – had much higher within geo-group collaboration intensities in 2014 than predicted. Panel (a) and (b) also show that, by and large, countries from Central Africa and Northern Africa are mostly located at the left bottom corner. This means that most countries from Central and Northern African regions are less active in performing international collaborations within their geographical groups.

The middle two panels (c) and (d) of Fig. 2 present the results regarding international collaborations within Africa (including all five geographical groups). At the earlier stage, shown in panel (c), most countries are located below the diagonal line, indicating that the actual international collaborations between African countries in 2007 were less intensive than predicted on the basis of the 2000 network. In other words, the collaborations between African countries until 2007 did not develop as much as predicted according to the network seven years ago. The three exceptional countries are Lesotho (no. 38), Swaziland (no. 41) and Comoros (no. 18), which developed higher within Africa collaboration links than predicted. This may be due to the fact that all three countries are small. The first two are completely and almost completely landlocked by South Africa, while the third one is a group of small islands. At the later stage, in panel (d), more countries moved up above the diagonal line. In particular Togo (no. 16), Guinea (no. 7) and Burundi (no. 17) were the ones that presented obviously higher within-Africa collaborations in 2014 than predicted. An interesting observation from panel (c) and (d) is that, in either temporal case, South Africa (no. 40) and Egypt (no. 51) were at the left bottom. That is, these two countries had rather low predicted within-Africa collaboration links, and moreover, their actual collaboration intensity became even lower in reality than predicted.

In contrast to the within geo-group and within Africa collaborations described above, the last two panels (e) and (f) show that the intensity of outside-Africa collaborations was significantly higher than predicted for the majority of African countries. In particular countries such as South Africa (no. 40), Egypt (no. 51), Algeria (no. 50), Morocco (no. 54) and Tunisia (no. 53) had collaboration intensities with non-African countries that were even higher than those predicted, even though the predicted collaboration intensities were already high (around 70% predicted based on network in 2000 and 80% predicted based on network in 2007 versus around 90% in the actual networks of both 2007 and 2014).

Figure 2: Comparison between predicted and actual collaboration intensities

Table 4 summarizes the predicted and actual collaboration intensities by geographical group. At the aggregated level, the actual non-African collaboration intensity in 2007 (74.2%) was higher than the

predicted value (58.0%). Northern Africa has the highest intensity in both actual (92.0%) and predicted (63.6%) networks. Accordingly, the international collaborations within Africa were lower than predicted, being at 25.8% in the actual network in contrast to the predicted 42%.

At the later stage, with regard to collaborations within African and outside of Africa, the actual intensities are close to the predicted values. At the aggregated level, collaboration intensity within Africa is 29.3%, and collaboration intensity outside Africa is 70.7% in 2014, which were almost the same as the prediction based on the network structure in 2007. Here, again, we find that Northern Africa is the region that is most oriented towards countries outside of Africa and that diverges most from the predicted values. In general, we observe that Northern Africa is only weakly integrating in Africa as a whole.

Comparing the earlier and later time periods, we can also observe a number of remarkable differences between the African regions. In terms of international collaborations within the same geo-group, the intensity in Western Africa has increased from 12.7% in 2007 to 21.3% in 2014, reflecting a strong regionalization process. A different evolution took place in Southern Africa, where the intensity of collaborations within geo-group has decreased, from 19.9% in 2007 to 11.7% in 2014. The regionalization process stays low in Northern Africa, below 6% in both stages.

Collaborations with outside Africa, however, present a different picture. In Northern Africa, the outside-Africa collaboration intensity stayed high, above 90% in both time periods. This is partially due to geographic and cultural ties with Saudi Arabia, which alone accounts for 14% of non-African collaborations with Northern Africa in 2014, but less than 1.5% for the other regions. Southern Africa is the only region which increased outside-Africa collaboration intensities, but reduced collaboration intensities within Africa. This indicates that scientists in Southern Africa have been working with more and more partners from outside of Africa. This is most likely due to the influence of South Africa.

Table 4: Summary of international collaboration intensity by geographical group

	Western Africa	Eastern Africa	Southern Africa	Central Africa	Northern Africa	All Africa
Predicted _based on data 2000						
With Africa Geo-group	13.7%	15.2%	4.5%	5.6%	10.3%	36.2%
Within Africa	45.1%	45.0%	42.2%	46.7%	36.4%	42.0%
Outside Africa	54.9%	55.0%	57.8%	53.3%	63.6%	58.0%
Actual collaboration intensity_2007						
With Africa Geo-group	12.7%	14.9%	19.9%	5.2%	5.8%	12.4%
Within Africa	29.9%	25.8%	37.8%	21.8%	8.0%	25.8%
Outside Africa	70.1%	74.2%	62.2%	78.2%	92.0%	74.2%
Predicted _based on data 2007						
With Africa Geo-group	9.6%	12.6%	4.7%	4.8%	7.2%	46.4%
Within Africa	37.2%	36.0%	33.0%	37.5%	26.5%	31.6%
Outside Africa	62.8%	64.0%	67.0%	62.5%	73.5%	68.4%
Actual collaboration intensity_2014						
With Africa Geo-group	21.3%	12.1%	11.7%	8.8%	5.9%	13.7%
Within Africa	38.9%	26.7%	27.7%	30.2%	9.1%	29.3%
Outside Africa	61.1%	73.3%	72.3%	69.8%	90.9%	70.7%

By relating the collaboration intensity with national scientific capacity (proxied by total publications in each country), we further observe from Figure A1 in the Appendix that scientifically stronger countries (i.e. countries denoted by large circles in Figure A1) tend to have a higher international collaboration intensity with non-African countries. On the contrary, African countries with lower scientific capacity (i.e. countries represented by small circles in Figure A1) tend to have relatively more collaborations with African countries. With the exception of Libya, the North African countries have relatively higher publication counts and also have higher intensities of international collaborations with partners outside of Africa. South Africa is the only country in the Southern Africa group which is scientifically strong, and also has a very high intensity of non-African collaborations (around 90 per cent). This shows that the most capable African countries (i.e. with higher publication output) have a higher tendency to collaborate with researchers outside of Africa. Less capable countries – with lower scientific capacity – are more active in shaping the network within Africa. As pointed out by Powell et al. (2005), “resource-rich participants are more capable of altering their positions by reconfiguring their networks.” In the case of African countries, our results are in line with this statement.

Another interesting observation from Figure 2 and Table 4 is that the emergent collaboration links – reflected by the difference between predicted and actual collaboration intensities – were more pronounced in the earlier stage (between 2000 and 2007) than in the later stage (between 2007 and 2014). In accordance with the numbers of nodes and links reported in Table 1, this illustrates that the international collaboration network structure became more stable over the years and the growth of new nodes and links has slowed down. For instance, 32 per cent new nodes were added between 2000 and 2007 (from 100 to 132), while 20 per cent new ones were added between 2007 and 2014 (from 132 to 159). The number of links grew 134 per cent between 2000 and 2007 (from 1756 to 4119), while it grew 95 per cent between 2007 and 2014 (from 4119 to 8033). Hence African countries were more active in reshaping the network in particular before 2007.

6 Discussion and conclusions

This paper compares Africa’s scientific collaboration network with the predicted scenarios and explores the emergence of new scientific cooperative partners and their locations. On the methodological level, we have shown how link prediction techniques can predict the majority of links and link weights reasonably well. By focusing on deviations between predicted and realized collaboration intensities, we can assess the impact of exogenous factors to the collaboration network’s evolution.

Our results show that African countries tend to collaborate with countries outside of Africa more than could be expected based on the intrinsic features of the earlier network, with the exception of Western Africa in 2007–2014. In particular, capable African countries (such as South Africa, Egypt, Morocco and Tunisia) greatly increased their scientific cooperation with countries outside of Africa compared to the predictions, despite their already high non-Africa collaboration intensity. The emergence of such links shows that non-African countries are still playing an important role in Africa’s scientific production.

In conducting scientific research, the highly visible and productive researchers are often believed to be the ones who are able to select partners who can further help improve their recognition (Wagner & Leydesdorff, 2005). In the case of Africa, our results show that, compared with those from

scientifically weak countries, scientists from Egypt and South Africa have a higher propensity to work with scientific partners from non-African economies. It should be noted, however, this is a generalized phenomenon at the country level, while a much greater variability exists at the individual level. Besides that, geography, history, and language are important factors in explaining variations in collaboration (Adams et al., 2014).

While most African countries exhibited lower within-Africa collaboration intensity than predicted, a few countries collaborated more with other African countries than expected. However, this only happened in countries that are relatively scientifically weak, such as Lesotho or Togo. This may be due to the barrier of “institutional mismatch” as explained by Georghiou (1998). The structures and priorities of scientific research in those African countries can be very different from those of countries outside of Africa, thus they intend to collaborate with countries with a similar economic or institutional background. It also seems likely that research priorities in scientifically weaker African countries are less compatible with those of international journals that are indexed in WoS or Scopus (Larivière, 2014).

Comparing different temporal stages, we find that the intensity of collaboration with non-African countries has decreased between the earlier stage (between 2000 and 2007) and the later stage (between 2007 and 2014). In the later stage, the predicted scenario was close to the actual network. This seems to be due to the fact that the growth of new nodes and links in the network has been slowing down and the number of unexpected collaborations beyond the existing network has reduced. As a result, the collaboration structure has become more stable over years. All geographical groups, except Southern Africa, have shown a regionalization process in the later stage. Western Africa is the group where collaborations within its own geo-group have increased most remarkably.

A clear limitation of the present work is that it is based on Scopus, a database that offers only limited coverage of non-Western journals and proceedings. This may distort the picture and give, for instance, the impression that non-English-speaking countries are scientifically weaker than they are in reality. Our study shares this limitation with several other studies of African research. Future research could focus on integrating other data sources that provide better coverage of local and non-English publication outlets.

As suggested by a reviewer, the link prediction approach could be further improved by taking into account all intervening years to arrive at a ‘final’ prediction. Another way in which future research might move beyond the present study is in exploring networks with fractionalized link weights (Gauffriau, Larsen, Maye, Roulin-Perriard, & von Ins, 2008; Perianes-Rodriguez, Waltman, & van Eck, 2016). The downside of our full counting approach is that a small number of publications written by co-authors from many countries may have a significant effect on the resulting network. This downside can be avoided by fractional counting.

Finally, it is worth noting that our focus on international collaborations implies that intra-national collaborations are not studied. It could be the topic of a follow-up study to do a similar analysis that takes intra-national collaborations into account and characterize the resulting picture using the model of Moed and Halevi (2014).

Acknowledgements

We thank the reviewers for their comments, which helped to improve the paper. Research by RG has been made possible by the financial support of the Flemish government to ECOOM.

7 References

- Adams, J., Gurney, K., Hook, D., & Leydesdorff, L. (2014). International collaboration clusters in Africa. *Scientometrics*, 98(1), 547–556. <http://doi.org/10.1007/s11192-013-1060-2>
- Adams, J., King, C., & Hook, D. (2010). *Global research report. Africa*. Leeds.
- Antonellis, I., Molina, H., & Chang, C. (2008). Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment*, 1(1), 408–421. <http://doi.org/10.1145/1453856.1453903>
- AOSTI. (2014). *Assessment of Scientific Production in the African Union, 2005–2010*.
- Arvanitis, R., Waast, R., & Gaillard, J. (2000). Science in Africa: A bibliometric panorama using PASCAL database. *Scientometrics*. <http://doi.org/10.1023/a:1005615816165>
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509.
- Boshoff, N. (2009). Neo-colonialism and research collaboration in Central Africa. *Scientometrics*, 81(2), 413–434. <http://doi.org/10.1007/s11192-008-2211-8>
- Boshoff, N. (2010). South-South research collaboration of countries in the Southern African Development Community (SADC). *Scientometrics*, 84(2), 481–503. <http://doi.org/10.1007/s11192-009-0120-0>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <http://doi.org/10.1023/A:1010933404324>
- Chataway, J., Smith, J., & Wield, D. (2005). Partnerships for Building Science and Technology Capacity in Africa. In *Paper prepared for the Africa–Canada–UK Exploration: Building AfricaCanadaUK Exploration building Science and Technology Capacity with African Partners*. London UK.
- Confraria, H., Godinho, M. M., & Wang, L. (2017). Determinants of citation impact: A comparative analysis of the Global South versus the Global North. *Research Policy*, 46(1), 265–279. <http://doi.org/10.1016/j.respol.2016.11.004>
- De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 292–306.
- Gaufriau, M., Larsen, P., Maye, I., Roulin-Perriard, A., & von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, 77(1), 147–176.
- Georghiou, L. (1998). Global cooperation in research. *Research Policy*, 27, 611–626. [http://doi.org/10.1016/S0048-7333\(98\)00054-7](http://doi.org/10.1016/S0048-7333(98)00054-7)
- Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2), 1461–1473. <http://doi.org/10.1007/s11192-013-1228-9>

- Jacobs, D., & Pichappan, P. (2006). Research Collaborations and Scientific productivity among the Research Universities in South Africa. In *In International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, Nancy (France), May 10-12, 2006*.
- Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining SE - KDD '02* (pp. 538–543). New York, NY, USA: ACM. <http://doi.org/10.1145/775047.775126>
- Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43. <http://doi.org/10.1007/bf02289026>
- Kreimer, P., & Meyer, J.-B. (2008). Equality in the networks? Some are more equal than others. *Universities as Centres of Research and Knowledge Creation: An Endangered Species*, 121–134.
- Langville, A., & Meyer, C. (2005). A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Review*, 47(1), 135–161. <http://doi.org/10.1137/s0036144503424786>
- Larivière, V. (2014). The importance of national journals. Retrieved May 12, 2017, from <http://www.universityaffairs.ca/opinion/in-my-opinion/importance-national-journals/>
- Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019–1031. <http://doi.org/10.1002/asi>
- Mêgnigbêto, E. (2013). International collaboration in scientific publishing: the case of West Africa (2001–2010). *Scientometrics*, 96(3), 761–783. <http://doi.org/10.1007/s11192-013-0963-2>
- Moed, H. F. (2016). Iran's scientific dominance and the emergence of South-East Asian countries as scientific collaborators in the Persian Gulf Region. *Scientometrics*, 108, 305–314. <http://doi.org/10.1007/s11192-016-1946-x>
- Moed, H. F., & Halevi, G. (2014). Tracking scientific development and collaborations – The case of 25 Asian countries. *Research Trends*, 38.
- Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213–228. <http://doi.org/10.1007/s11192-015-1765-5>
- Murata, T., & Moriyasu, S. (2007). Link Prediction of Social Networks Based on Weighted Proximity Measures. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 85–88). Washington, DC, USA: IEEE Computer Society. <http://doi.org/10.1109/wi.2007.71>
- Ocholla, D. N., Mostert, J., & Rotich, D. C. (2016). Visibility of University of Zululand and Moi University researchers in Web of Science and Scopus from 2003 to 2013. *African Journal of Library Archives and Information Science*, 26(1), 3–15.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Perianes-Rodriguez, A., Waltman, L., & van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4), 1178–1195. <http://doi.org/10.1016/j.joi.2016.10.006>

- Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. *American Journal of Sociology*, 110(4), 1132–1205. <http://doi.org/10.1086/421508>
- Rattigan, M. J., & Jensen, D. (2005). The case for anomalous link discovery. *ACM SIGKDD Explorations Newsletter*, 7(2), 41–47. <http://doi.org/10.1145/1117454.1117460>
- Schubert, T., & Sooryamoorthy, R. (2010). Can the centre-periphery model explain patterns of international scientific collaboration among threshold and industrialised countries? The case of South Africa and Germany. *Scientometrics*, 83(1), 181–203. <http://doi.org/10.1007/s11192-009-0074-2>
- Sooryamoorthy, R. (2009). Collaboration and publication: How collaborative are scientists in South Africa? *Scientometrics*, 80(2), 419–439. <http://doi.org/10.1007/s11192-008-2074-z>
- Toivanen, H., & Ponomariov, B. (2011). African regional innovation systems: Bibliometric analysis of research collaboration patterns 2005-2009. *Scientometrics*, 88(2), 471–493. <http://doi.org/10.1007/s11192-011-0390-1>
- Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10), 1608–1618. <http://doi.org/10.1016/j.respol.2005.08.002>
- Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, 8(2), 295–309. <http://doi.org/10.1016/j.joi.2014.01.008>
- Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(4), 623–630. <http://doi.org/10.1140/epjb/e2009-00335-8>

Appendix:

Table A1: Group list of African countries

Western Africa	Eastern Africa	Southern Africa	Central Africa	Northern Africa
1. Benin	17. Burundi	37. Botswana	42. Angola	50. Algeria
2. Burkina Faso	18. Comoros	38. Lesotho	43. Cameroon	51. Egypt
3. Cape Verde	19. Djibouti	39. Namibia	44. Centre Afr. Rep.	52. Libya
4. Cote d'Ivoire	20. Eritrea	40. South Africa	45. Chad	53. Morocco
5. Gambia	21. Ethiopia	41. Swaziland	46. Congo	54. Tunisia
6. Ghana	22. Kenya		47. Equatorial Guinea	
7. Guinea	23. Madagascar		48. Gabon	
8. Guinea-Bissau	24. Malawi		49. Sao Tome & Principe	
9. Liberia	25. Mauritius			
10. Mali	26. Mayotte			
11. Mauritania	27. Mozambique			
12. Niger	28. Réunion			
13. Nigeria	29. Rwanda			
14. Senegal	30. Seychelles			
15. Sierra Leone	31. Somalia			
16. Togo	32. Sudan			
	33. Tanzania			
	34. Uganda			
	35. Zambia			
	36. Zimbabwe			

Source: from The United Nations Statistics Division. <https://unstats.un.org/unsd/methodology/m49/>

Note: 1) "Democratic Republic of the Congo" is combined with Congo. 2) "Sudan" is combined with "South Sudan" in the Eastern Africa group. 3) "United Republic of Tanzania" is called "Tanzania" in our list. 4) "Saint Helena" and "Western Sahara" are not included.

Figure A1: Correlation of predicted and actual collaboration intensity (relating to scientific capacity)
 Note: 1) X-axis denotes the average of predicted values (based on the data in 2000 and 2007), and y-axis represents the average of actual collaboration intensity (based on the reality in 2007 and 2014). 2) The size of circles reflects the average volume of scientific publications (mean of publications in 2000, 2007 and 2014).