Suggesting Collaborations between African and South-Asian Countries

Raf Guns[*] and Ronald Rousseau[**]

[*] *raf.guns@ua.ac.be*
IOIW, University of Antwerp, Venusstraat 35, B-2000 Antwerpen (Belgium)

[**] *ronald.rousseau@khbo.be*
IOIW, University of Antwerp, Venusstraat 35, B-2000 Antwerpen (Belgium)
KHBO (Association K.U Leuven), Faculty of Engineering Technology, Zeedijk 101, B-8400 Oostende (Belgium)
KU Leuven, Department of Mathematics, Celestijnenlaan 200B, B-3001 Heverlee (Belgium)

Introduction

We study collaboration at the country level between African and South-Asian countries for research on two specific topics (*malaria* and *tuberculosis*). The following countries are considered:

- all African countries;
- all countries in the Middle East, except for Israel and Turkey (considered to be more European oriented);
- countries in South-Asia, that is, all Asian countries excluding countries that belonged to the former Soviet Republic, Mongolia, China, North and South Korea, Taiwan and Japan.

After examining the current structure of the collaboration networks, we look for (as of yet unrealized) opportunities for future collaboration, using a link prediction approach. In this way, we can make concrete recommendations for collaboration at the level of institutes. Note that not all included countries can be considered developing nations in a strict sense.

Data and methods

All data were collected from Thomson Reuters' Web of Science (WoS) on January 31, 2012. We searched for all publications published in the five-year period 2007-2011 with at least one address in one of the above mentioned countries. Restricting this set to the two topics yielded 7,762 publications on malaria and 7,360 on tuberculosis.

For each publication, the country of each author's (primary) affiliation was recorded. Some country names had to be normalized: Burma and Myanmar were merged and the Republic of the Congo (Congo-Brazzaville) had to be distinguished from the Democratic Republic of the Congo (Congo-Kinshasa) in some cases. All countries that co-occur on a publication are then linked, resulting in a weighted network of collaboration between countries (per topic). Link weights denote the number of publications with authors from the two countries. Because our analysis is on the level of countries rather than individuals, a publication with five authors from country A and three from country B is treated the same as a publication with one author from A and one from B.

Some publications in our data have co-authors from countries outside the set specified above. Therefore, we created two networks for each topic: a network including these external countries (the full network) and a network excluding them (the reduced network). The characteristics of these collaboration networks are discussed in the next section.

Collaboration network structure

Table 4 provides basic descriptive statistics per network. Western countries such as the USA and England dominate the full networks, even though South Africa is also an important node in the full tuberculosis network. The reduced malaria network is dominated by Kenya, Tanzania, South Africa and Thailand, while the reduced tuberculosis network is dominated by India and South Africa.

Table 4. Network statistics per topic.

Network		Nodes	Links	Density
mala-	full	146	2,333	0.220
ria	reduced	75	647	0.233
tuber-	full	144	2,092	0.203
culosis	reduced	64	395	0.196

Link prediction for recommendation

Since we are interested in opportunities for future collaboration, we focus on countries that do not yet collaborate according to our data. There are many possible methods for determining which future collaborations are the most promising. Here, we only use the information that is already present in the country collaboration network. More specifically, we take a link prediction approach, trying to determine a likelihood score W for each node pair on the basis of the current network. Singling out those pairs that are currently unlinked and sorting them in decreasing order of W yields a list of the most promising future collaborations.

A formula that results in a likelihood score *W* is called a predictor. We use three predictors that had good performance in previous research (Guns, 2011; Liben-Nowell & Kleinberg, 2007).

Weighted Katz predictor

The weighted Katz measure can best be described in the context of a multigraph (a graph allowed to have multiple links between two nodes). Let A denote the (full) adjacency matrix of the multigraph M. The element a_{ij} is equal to the number of links between v_i and v_j or 0 if no link is present. Each element $a_{ij}^{(k)}$ of A^k (the k-th power of A) has a value equal to the number of walks in M with length k from v_i to v_j. The weighted Katz predictor is then defined as (Katz, 1953):

$$W(v_i, v_j) = \sum_{l=1}^{\infty} \beta^k a_{ij}^{(k)}$$

where β is a parameter between 0 and 1. This parameter represents the "probability of effectiveness of a single link". Thus, each path with length k has a probability β^k of effectiveness. We have taken $\beta = 0.001$, based on previous experiences.

Rooted PageRank

The intuition behind rooted PageRank (Liben-Nowell & Kleinberg, 2007) is best explained from the perspective of a random walker. The random walker starts at a fixed node v, called the root node. At each step, it moves along a link to a neighbour of the current node. Contrary to ordinary PageRank, rooted PageRank does not allow random 'teleportation' but only allows teleportation back to the root node . This form of teleportation occurs with probability $1 - \alpha$. High α values tend to favour the well-connected nodes in the network (with high classic PageRank scores), especially in relatively small networks such as ours. We therefore set $\alpha = 0.4$.

SimRank

SimRank is a measure of the similarity of two nodes in a network (Jeh & Widom, 2002). The SimRank thesis is: *Objects that link to similar objects are similar themselves*. The starting point is the assumption that an object is maximally similar to itself: $Sim(a, a) = 1$. One can then calculate the SimRank score of each node pair iteratively, using the SimRank formula:

$$W(u, v) = \frac{c}{|N_u| \cdot |N_v|} \sum_{p \in N_u} \sum_{q \in N_v} W(p, q)$$

where $|N_v|$ denotes the number of nodes adjacent to *v* and c is a constant.

Recommended collaborations

The three predictors yield different results because their underlying philosophies are also different. Hence, predictions that rank high for all three predictors are all the more interesting. We will focus on these predictions.

For malaria research our main suggestions are: Madagascar–India, Kenya–Malaysia and India–Gambia. For tuberculosis research our main proposals are: India–Ethiopia, South Africa–Bangladesh and Mozambique–India. As India has relatively few research collaborations with other countries (Glänzel & Gupta, 2008), it is not surprising that collaborations with India are among the strong suggestions.

African and South-Asian countries usually have just one (or two) top institute(s) working on a certain topic. Hence our country-country suggestions for collaboration can also be interpreted as suggestions for institutional collaboration as shown in Table 2.

Table 2. Collaboration suggestions for malaria

Institute A	**Institute B**
Institut Pasteur de Madagascar	National Institute of Malaria Research (India)
Kenya Medical Research Institute	University of Malaysia Sarawak (UNIMAS) or University Sains Malaysia
National Institute of Malaria Research (India)	MRC Labs (Medial Research Council Gambia)

Conclusion

Link prediction techniques can be used to suggest research collaborations.

References

Glänzel, W. & Gupta, B.M. (2008). Science in India. A Bibliometric Study of National Research Performance in 1991-2006. *ISSI Newsletter*, 4 (3), 42-48.

Guns, R. (2011). Bipartite Networks for Link Prediction: Can they improve Prediction Performance? In *Proceedings of the ISSI 2011 Conference* (pp. 249–260). Leiden: Leiden University.

Jeh, G., & Widom, J. (2002). SimRank: A Measure of Structural-Context Similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference* (pp. 538–543). New York: ACM.

Katz, L. (1953). A New Status Index Derived from Sociometric Analysis. *Psychometrika*, 18 (1), 39-43.

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.