Recommending research collaborations using link prediction and random forest classifiers

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

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2), 1461–1473.

Recommending research collaborations using link prediction and random forest classifiers

Raf Guns[1] and Ronald Rousseau[2]

[1] *raf.guns@uantwerpen.be*
University of Antwerp, Institute for Education and Information Sciences, IBW,
Venusstraat 35, B-2000 Antwerp, Belgium

[2] *ronald.rousseau@uantwerpen.be*
University of Antwerp, Institute for Education and Information Sciences, IBW,
Venusstraat 35, B-2000 Antwerp, Belgium

KU Leuven, B-3000 Leuven, Belgium

Abstract

We introduce a method to predict or recommend high-potential future (i.e., not yet realized) collaborations. The proposed method is based on a combination of link prediction and machine learning techniques. First, a weighted co-authorship network is constructed. We calculate scores for each node pair according to different measures called predictors. The resulting scores can be interpreted as indicative of the likelihood of future linkage for the given node pair. To determine the relative merit of each predictor, we train a random forest classifier on older data. The same classifier can then generate predictions for newer data. The top predictions are treated as recommendations for future collaboration.

We apply the technique to research collaborations between cities in Africa, the Middle East and South-Asia, focusing on the topics of malaria and tuberculosis. Results show that the method yields accurate recommendations. Moreover, the method can be used to determine the relative strengths of each predictor.

Introduction

Research collaboration is an important topic in informetrics. Collaboration has the potential of saving costs and diffusing insights and ideas between partners. Hence, the advantages of collaboration are very attractive to institutes in those regions or countries that do not yet belong to the 'rich and famous' in science. While it may seem most attractive to collaborate with wealthier regions (Schubert & Sooryamoorthy, 2010), there are several advantages when collaborating among developing nations, such as the establishment of local centres of excellence and a greater awareness among partners of the needs and problems common to developing nations (Boshoff, 2010). Yet it is not always obvious which partners one should collaborate with. Using recommendation techniques is a possible way to approach this problem (e.g., Yang & Jin, 2006). The current article proposes a new practical method to generate collaboration recommendations for policy makers and university strategists.

In this article, we study research collaboration between cities in Africa, the Middle East, and South-Asia. Co-authorship networks are constructed among these cities within the research fields of malaria and tuberculosis during three consecutive, five-year time periods: 1997–2001, 2002–2006, and 2007–2011. Our aim is to develop a methodology for recommending potentially fruitful collaborations, using link prediction and machine learning. The method generates recommendations by 'learning' from the first two time periods. We evaluate the

quality of the generated recommendations by comparing them with the actual collaborations in the third period.

In the next section, we discuss how the data has been collected. Subsequently, we discuss the extraction of the collaboration networks and explain our link prediction and machine learning approach. The Results section highlights the recommended collaborations and their quality. The final section contains the conclusions.

This article is a reworked and extended version of a paper that was presented at the ISSI 2013 conference (Guns & Rousseau, 2013).

Data

Cities located in the following countries (referred to as the target countries) are included if they have contributions in the field under study:

- 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 belong to the former Soviet Republic, Mongolia, China, North and South Korea, Taiwan and Japan.

We restrict ourselves to two topics: the diseases *malaria* and *tuberculosis*. These are topics that are not entirely dominated by Western countries on the one hand, and not too specific to a certain country or region (The STIMULATE-6 Group, 2007) on the other.

The data were collected from Thomson Reuters' Web of Science (WoS) on October 26 and November 21, 2012. We searched for all publications published in the three five-year periods (1997–2001, 2002–2006, 2007–2011) with at least one address in one of the target countries. These sets were then restricted to the two topics. Results are summarized in Table 1.

Table 1. Numbers of publications for each topic and period

Topic	Number of publications		
	1997–2001	2002–2006	2007–2011
malaria	2,622	4,671	7,901
tuberculosis	2,369	3,830	7,832

Table 2. Number of cities in the data

Topic	Number of cities (African and South-Asian / other)		
	1997–2001	2002–2006	2007–2011
malaria	400 / 361	601 / 587	904 / 883
tuberculosis	351 / 270	482 / 468	831 / 777

Methods

Network construction

After exporting the search results from the WoS, we extracted a weighted network of co-authorship between cities as follows (for both topics and for each time period). For each publication, the city of each author's (primary) affiliation was recorded. A script was written

to extract the city automatically. Because of the large variety of address formats and inconsistencies in the data, all results were manually checked and corrected where necessary. Table 2 summarizes the results.

Subsequently a network was created whose node set consists of all cities encountered. All cities that co-occur on a single publication are then linked in the network. The weight of the link between cities A and B is the number of publications with authors from A and B. Because our analysis is on the level of cities rather than individuals, we have not taken into account the number of authors from a city on a single publication. For instance, a publication with five authors from city A and three from city 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 cities outside the set of target countries (see 'other' in Table 2). Therefore, we decided to create two networks for each topic: a network including these external cities – the full network – and a network excluding them – the restricted network. In total, this procedure led to twelve different networks: a full and a restricted network for each of the two topics, and this for each of the three periods. First we describe the link prediction techniques we use, followed by the machine learning technique.

Link prediction techniques

Since we are interested in opportunities for future collaboration, we focus on cities from the target countries that do not yet collaborate in a given time period. There are many possible methods for determining which future collaborations are the most promising. Here, we focus on the information that is already present in the city collaboration network, without relying on any other data source. We start from the assumption that a collaboration should be recommended if (a) the two cities do not yet collaborate, and (b) the two cities are similar or related. To determine the similarity or relatedness of cities, we take a link prediction approach. We try to determine a relatedness score W for each node pair on the basis of the current network. Singling out those pairs that are currently unlinked (condition a) and sorting them in decreasing order of W (condition b) yields a list of the most promising future collaborations.

A formula that results in a relatedness score W is called a predictor. We use the following predictors (Guns, 2011, 2012): common neighbours, cosine, Adamic/Adar, weighted graph distance, weighted Katz, weighted rooted PageRank, and weighted SimRank. Taken together, these predictors represent all of the most important approaches that were studied in the seminal paper of Liben-Nowell and Kleinberg (2007): raw co-occurrence (common neighbours), normalized co-occurrence (cosine, Adamic/Adar), distance-based approaches (weighted graph distance, weighted Katz), and approaches based on random walks (weighted rooted PageRank, weighted SimRank). Guns (2012) showed that the weighted versions of the last four predictors outperform the unweighted counterparts; for this reason we only include the weighted versions.
We will now discuss each of these in turn.

Common neighbours. Common neighbours is defined as the number of neighbouring nodes that two nodes have in common:

$$W(u,v) = |N(u) \cap N(v)| \qquad (1)$$

where $N(v)$ denotes the set of neighbours of node v and $|S|$ denotes the number of elements in set S.

Jaccard. The Jaccard predictor is a normalization of common neighbours:

$$W(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{2}$$

Adamic/Adar. The Adamic/Adar predictor starts from the assumption that two nodes u and v having (many) low-degree neighbours in common indicates a stronger relatedness between u and v than having highly connected neighbours in common:

$$W(u,v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log N(z)} \tag{3}$$

This measure was initially defined in the context of social networks on the Web (Adamic & Adar, 2003).

Weighted rooted PageRank. The next two predictors are inspired by Google's PageRank (and hence indirectly by the Pinski-Narin citation influence methodology (1976)). 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, the walker 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 v. This form of teleportation occurs with probability $1 - \alpha$ (where $0 < \alpha < 1$). 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. On the other hand, setting α too low reduces the advantage of a PageRank-like predictor. We use link weights, such that a link with a higher weight will be more likely to be traversed than a link with a lower weight.

Essentially, rooted PageRank is a specific form of so-called personalized PageRank (Langville & Meyer, 2005). The resulting scores can be interpreted as a measure of each node's relatedness to the root node. The highest scoring node is typically the root node itself.

Weighted SimRank. 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*. Note the recursive nature of the thesis, – to assess the similarity of a node pair, we need to have an estimate of the similarity of the nodes that they link to. The starting point of a SimRank computation is the assumption that an object is maximally similar to itself: $W(a,a) = 1$. One can then calculate the SimRank score of each node pair iteratively, until the changes drop below a given threshold value. The basic SimRank formula is:

$$W(u,v) = \frac{c}{|N_u| \cdot |N_v|} \sum_{p \in N_u} \sum_{q \in N_v} W(p,q) \tag{4}$$

In case of isolate nodes, the above formula would lead to a division by zero, which can be avoided by adding 1 to the denominator. Since our data contains no isolates, this is not necessary. Since we want to take link weights into account, we extend equation (4) as follows (Antonellis et al., 2008):

$$W(u,v) = \frac{c}{\sum_{p \in N_u} w(u,p) \cdot \sum_{q \in N_v} w(v,q)} \sum_{p \in N_u} \sum_{q \in N_v} W(p,q) \cdot w(u,p) \cdot w(v,q) \quad (5)$$

where $w(x,y)$ denotes weight of the link between x and y. In (4) and (5), c $(0 < c < 1)$ is the 'decay factor' that determines how quickly similarities decrease. If, for example, cities x and y both collaborate with z, then c determines the certainty with which we can state that x and y are similar. Lower c values also result in lower values for $W(x,y)$.

Weighted graph distance. One can hypothesize that the longer the distance between two nodes, the less related they are. Since we have a weighted network, the question is how to define distance in this context. It has been proposed by several authors (Egghe & Rousseau, 2003; Newman, 2001) that the path length in a proximity-based weighted network can be defined by taking the inverse of each weight:

$$p(u,v) = \sum_{i=1}^{t} \frac{1}{w_i} \quad (6)$$

where w_i $(i = 1, \dots, t)$ denotes the weight of the ith link in the path. A possible downside of this approach is that it ignores the number of nodes one has to traverse: in some cases, a large number of intermediary nodes may have a negative effect on the relatedness of the path's endpoints, even if the link weights are high. For this reason, Opsahl, Agneessens, and Skvoretz (2010) propose the following generalization of path length in a weighted network:

$$p_\alpha(u,v) = \sum_{i=1}^{t} \frac{1}{w_i^\alpha} \quad (7)$$

where α is an extra parameter between 0 and 1. If $\alpha = 0$, (7) reduces to the path length in the corresponding unweighted network – i.e., only the number of intermediary nodes is taken into account. If $\alpha = 1$, (7) reduces to (6) – i.e., only the link weights are taken into account. Thus, setting $0 < \alpha < 1$ allows us to find a balance between these two extremes. Because we want to favour short paths over longer ones, we define the weighted graph predictor as

$$W(u,v) = \frac{1}{p_\alpha(u,v)} \quad (8)$$

Weighted Katz. Before we define the weighted Katz predictor (Katz, 1953) we explain the used terminology. A *walk* is a sequence of nodes $v_1, v_2, \dots, v_m$, such that each node pair v_i, v_{i+1} in the sequence is connected by a link. There are no further restrictions on walks. A *multigraph* is a graph allowed to have multiple links between two nodes. Different links between two nodes also constitute different walks, i.e. the number of walks $v_1, v_2, \dots, v_m$ in a multigraph is equal to $\prod_{i=1}^{m-1} N(v_i, v_{i+1})$, where $N(v_i, v_{i+1})$ denotes the number of links between v_i and v_{i+1}.

The weighted Katz predictor can best be described in the context of a multigraph. 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 (Wasserman & Faust, 1994, p. 159). The weighted Katz predictor is then defined as:

$$W(v_i, v_j) = \sum_{k=1}^{\infty} \beta^k a_{ij}^{(k)} \qquad (9)$$

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. As $0 < \beta < 1$ higher powers become smaller and smaller so that the influence of nodes further away decreases fast.

Random forest classifiers

Now we turn to the machine learning technique that is used to aggregate the results of the different predictors. Random forests were introduced by Breiman (2001) as a robust machine learning technique for classification and regression. A random forest is an ensemble of decision trees, where each tree is built starting from a bootstrap 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. By introducing randomness at both the data and the model level, random forests have been shown to yield accurate and robust results. A practical advantage of random forests is that the procedure automatically predicts the probability with which an item belongs to a certain class (in our case: the probability of a link for each node pair). This is not the case for, for instance, support vector machines (SVMs), which require a time-consuming cross-validation procedure (Platt, 1999) to obtain similar probability estimates.

For our purposes, we need to classify node pairs (potential links) into two groups: links and non-links. The method works as follows:

(1) We split the data into two time periods, such that we have an early network A_1 and a later network A_2.
(2) We choose a number of predictors (see above). For each node pair in A_1 we calculate its relatedness score according to each predictor.
(3) The random forest classifier is trained on the features (relatedness scores) from A_1 and the corresponding classification data (link or not) from A_2. Essentially, the classifier learns each predictor's relative strength in predicting which links will or will not occur in the next time period.
(4) For each potential link in A_2 we calculate its relatedness score according to each predictor.
(5) The trained classifier yields predictions on the basis of the features that were determined in the previous step.
(6) We treat the top n predictions as recommendations for collaboration.

Here, we will use the 1997–2001 networks as A_1 and the 2002–2006 networks as A_2. We use the *scikit-learn* (Pedregosa et al., 2011) random forest classifier with 500 trees.

Evaluation

Using the networks from 1997–2001 and 2002–2006 for training purposes, we produce recommendations for the period following 2002–2006. In other words, we can use the actual data from the period 2007–2011 to assess the quality of our recommendations, i.e. to see whether or not the recommendations are realized.

Since we are interested in recommending high-potential collaborations, it makes sense to restrict our analysis to the top predictions. Concretely, we draw a list of the n unlinked node pairs with the highest score (we will test this for different values of n). These are considered as our recommendations. A recommendation is successful if it takes place in the evaluation period 2007–2011. Let s denote the number of successful recommendations and n the total number of recommendations. We then determine the success rate $SR = s/n$ as an indicator of recommendation quality. SR is essentially precision-at-n.

Suppose that the test network from the period 2007–2011 is a complete network, where every pair of nodes is connected. In that case, any prediction would lead to $SR = 1$. In general, if the test network is very dense, one would expect higher SR values. For this reason, we will also provide the expected success rate for a completely random prediction. This is equal to the ratio of the number of new links (links present in the test network but not the training network) and the number of possible new links (unlinked node pairs in the training network).

Results

Collaboration network structure

Before turning to the recommendation results, we briefly describe the structure of the collaboration networks that we obtained.

Using VOSviewer (Van Eck & Waltman, 2007, 2010) we obtained twelve visualizations of our data: one for each network. We provide one visualization for malaria (Figure 1) and one for tuberculosis (Figure 2). Colour versions of these figures as well as the other ten visualizations are available in the additional online material.

The malaria networks can be described as follows. In the full view (period 1997–2001) we can easily see a dense main cluster dominated by Oxford, London and Bangkok; Indian cities (New Delhi) have a peripheral position. During the period 2002–2006 the main cluster is dominated by London, Bangkok and Nairobi. Indian cities have moved closer to the main cluster. Finally, during the period 2007–2011 we have a strong main cluster, including Indian cities, and dominated by London and Oxford. When considering the restricted networks the 1997–2001 view is rather scattered with centres in Nairobi and Bangkok, with some Vietnamese cities between these two centres; Indian cities are situated far away from these clusters. During the period 2002–2006 the Vietnamese cluster has almost merged with the Thai one. Finally during the period 2007–2011 there is a clear African cluster (Nairobi, Dakar, Cape Town) and an Asian one (Bangkok, Mae Sot, New Delhi) as can be seen in Figure 1. Moreover an Iranian group of cities becomes visible on the periphery.

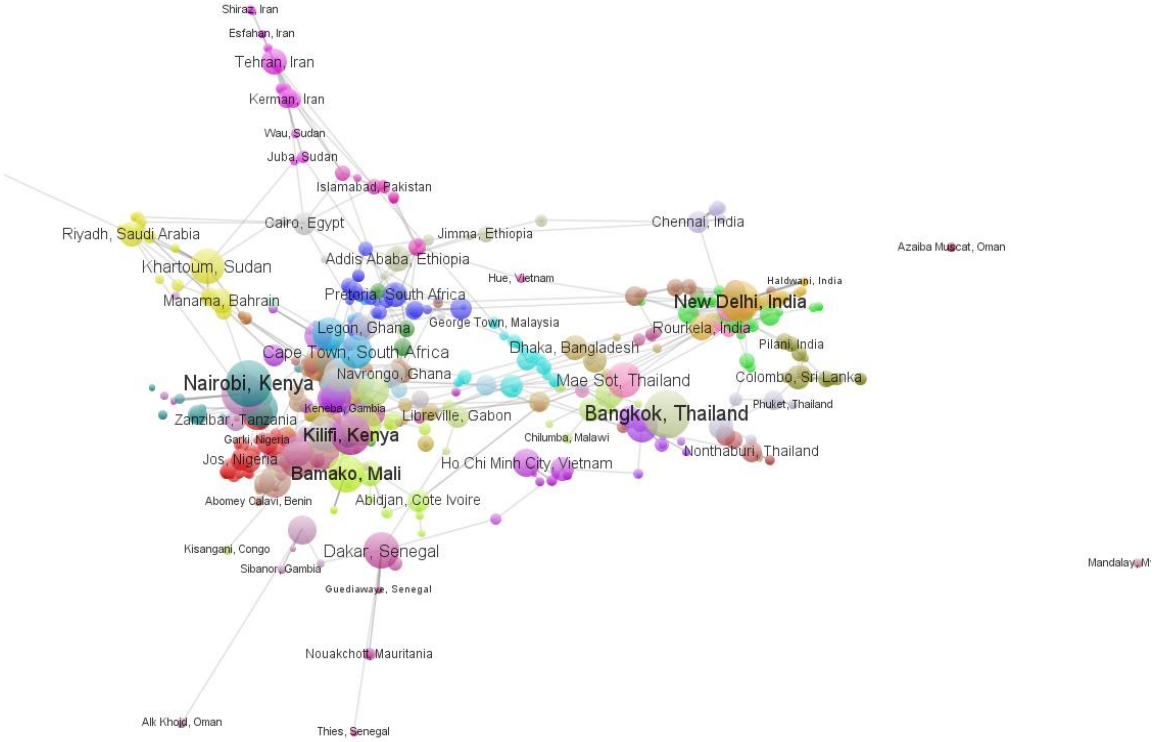

Figure 1. Collaboration network for malaria (restricted view, 2007–2011)

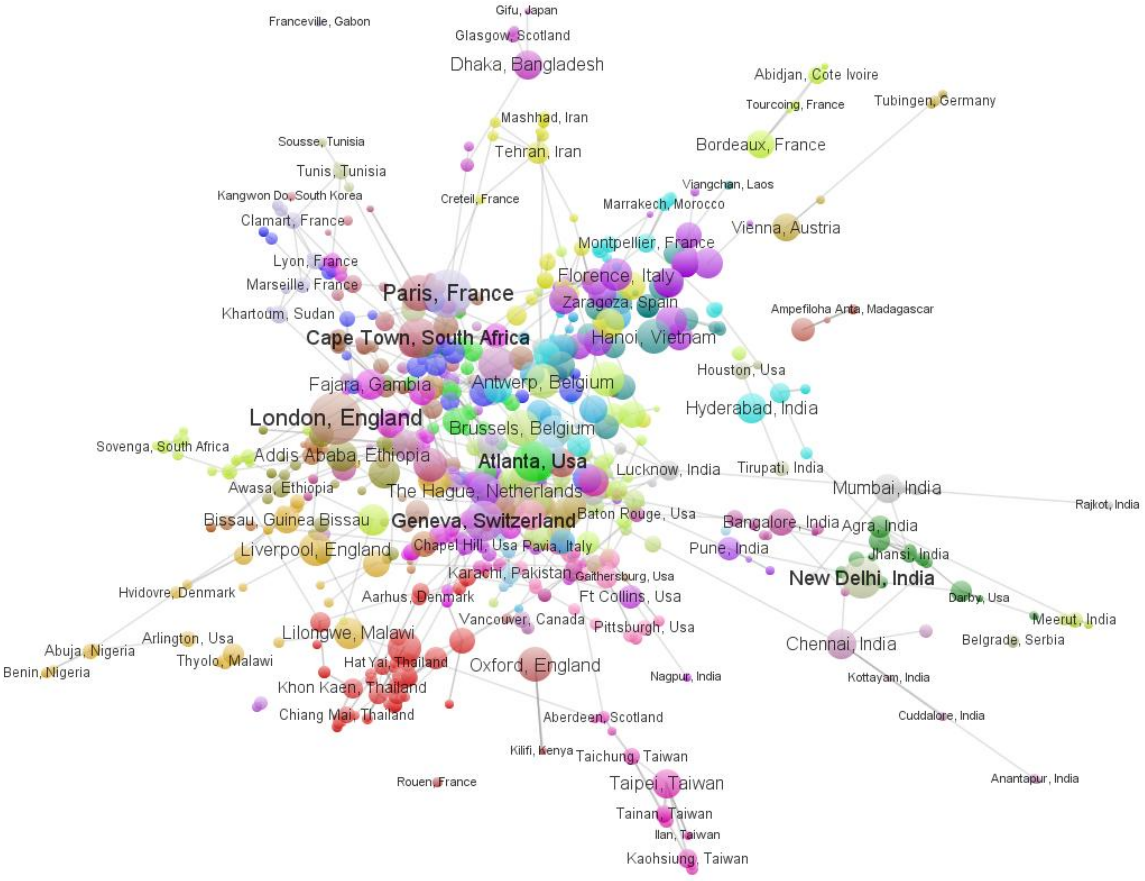

Figure 2. Collaboration network for tuberculosis (full view, 2002–2006)

As to the tuberculosis networks, the 1997–2001 full view shows a group of centres around London, Geneva, Atlanta, Paris and Johannesburg. These are situated rather close to one another. During the period 2002–2006 these groups have formed a main cluster where we see London, Geneva, Paris, New Delhi and Oxford; Dhaka (Bangladesh) is clearly visible above this main cluster, while Antwerp and Brussels (Belgium) are situated in the very centre of this figure (Figure 2). In the 2007–2011 view we again have several clusters situated close to one another. The largest, central, one contains London, Paris, Geneva, Cape Town, Kampala and Liverpool; close to this main cluster we have an Indian cluster around New Delhi and Chennai; we further have clusters around Taipei and around Tehran. The 1997–2001 restricted network contains several scattered clusters around the following centres: South-Africa (Cape Town, Johannesburg), Chennai-Pune, another Indian one around New Delhi, Bangalore and including Bangkok, and finally one around Addis Ababa (Ethiopia). The 2002–2006 view is very linear with centres around New Delhi, Hanoi, Bangkok and Cape Town (and other South African cities) including Dakar (Senegal). Finally the restricted 2007–2011 view contains a large cluster around Cape Town (and South African cities) and including Addis Ababa. Moreover we see an Indian cluster, a Thai one and an Iranian one on the periphery.

In summary, the following observations pertain to both topics. The full networks are mainly dominated by Western cities, although some larger African or Asian cities are also able to

occupy a central position. There appears to be at least a mild form of geographical bias – e.g., Asian cities mainly collaborating with other Asian cities – but the effect is modest: we also found several cases of intense international and intercontinental collaboration. Some countries, such as India and Iran, are more likely to form separate clusters. This observation corresponds with the results of Glänzel and Gupta (2008) who found that India has relatively few research collaborations with other countries.

Recommendations

We predicted collaborations between research institutes situated in different cities based on relatedness scores as explained above. The parameters for each predictor were set to the values shown in Table 3. These predictors were then applied to each of the four cases (malaria full, malaria restricted, tuberculosis full, and tuberculosis restricted). Applying the method whereby we train a random forest classifier on earlier predictions (namely all predictions obtained by the seven predictors used) yields the results that are summarized in Table 4.

Table 3. Values chosen for predictor parameters

Predictor and parameter	Value
Weighted graph distance: α	0.9
Weighted Katz: β	0.001
Weighted rooted PageRank: α	0.4
Weighted Simrank: c	0.3

Table 4. Recommendation success rates

Data set	Random success rate	Success rate				
		$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
Malaria (full)	0.02	1.00	1.00	0.95	0.82	0.77
Malaria (restricted)	0.02	0.60	0.60	0.80	0.68	0.67
Tuberculosis (full)	0.01	0.60	0.80	0.70	0.48	0.35
Tuberculosis (restricted)	0.01	0.60	0.50	0.45	0.52	0.37

In general, the success rate decreases as more recommendations are generated, although exceptions exist. The success rate of our method is clearly – by a factor of 35 to 80 times – better than randomized predictions. Predictions based on the full network are generally (but not always) better than those based on the restricted network. Since the former contains more information than the latter, this is not unexpected. Indeed, if two target cities collaborate with a Western city, they may eventually end up collaborating directly, but this can only be inferred from the full network. Comparing our results with the performance of individual predictors (Guns & Rousseau, 2013) illustrates the merits of the current method: only in one specific case (tuberculosis 2002–2006, restricted) does an individual predictor (weighted Katz) yield a higher success rate.

Table 5 provides the top 10 recommendations for the full and restricted malaria networks. It can be seen that most recommendations for African cities involve other African cities (and likewise for Asian ones), but some non-trivial cross-continental recommendations occur as

well. Successful predictions often involve South-African or Thai cities (e.g., Mae Sot, Johannesburg). Such cities could be called facilitator cities. They play a central role in weaving the fabric of international collaboration.

Table 5. Top 10 recommendations for malaria (full and restricted)

#	Malaria full	Malaria restricted
1	Banjul, Gambia – Bamako, Mali	Mae Sot, Thailand – Jakarta, Indonesia
2	Yaounde, Cameroon – Banjul, Gambia	Ifakara, Tanzania – Blantyre, Malawi
3	New Delhi, India – Bamako, Mali	Kilifi, Kenya – Addis Ababa, Ethiopia
4	Johannesburg, South Africa – Bamako, Mali	Kisumu, Kenya – Dakar, Senegal
5	Nairobi, Kenya – Antananarivo, Madagascar	Moshi, Tanzania – Kampala, Uganda
6	Johannesburg, South Africa – Dakar, Senegal	Ibadan, Nigeria – Calabar, Nigeria
7	Ouagadougou, Burkina Faso – Nairobi, Kenya	Durban, South Africa – Blantyre, Malawi
8	Yaounde, Cameroon – Accra, Ghana	Johannesburg, South Africa – Bangkok, Thailand
9	Dar Es Salaam, Tanzania – Cape Town, South Africa	Mae Sot, Thailand – Ifakara, Tanzania
10	Bamako, Mali – Accra, Ghana	Kuala Lumpur, Malaysia – Bangkok, Thailand

Since the random forest classifier aggregates the results of (in our case) seven predictors, it is interesting to explore the relative contribution of each predictor to the final results. We determine predictor importance using the so-called Gini importance measure (Breiman et al., 1984). At each split in a tree, one records the decrease in heterogeneity of predictions. The average of all decreases in the forest for a given predictor is its Gini importance. The higher the Gini importance, the more important the predictor.

Table 6 shows each predictor's contribution according to Gini importance. The predictors that are only based on neighbouring information contribute less than the topology-based predictors. Remarkably, the two predictors that contribute most, SimRank and rooted PageRank, do not yield very high success rates when used in isolation (Guns & Rousseau, 2013). At the same time, the individual predictor with the highest success rate (weighted Katz) contributes far less to the aggregated results. We hypothesize that this is due to the fact that the random forest classifier takes all predictions into account, whereas the high success rates for weighted Katz found by Guns and Rousseau (2013) were based on just twenty predictions. We also see that the most important predictors are those that incorporate link weights.

Table 6. Predictor importance, averaged over the four data sets

Predictor	Average Gini importance (± s.d.)
Weighted SimRank	0.264 (± 0.054)
Weighted rooted PageRank	0.241 (± 0.032)
Weighted graph distance	0.163 (± 0.016)
Weighted Katz	0.146 (± 0.017)

Adamic/Adar	0.085 (± 0.027)
Jaccard	0.072 (± 0.017)
CommonNeighbours	0.028 (± 0.014)

A downside of the method is that it tends to yield less interesting predictions: whereas single predictors regularly recommend collaborations between African and Asian cities, this is rare with our current method. In other words, an improvement of the method would try to strike a better balance between performance on the one hand and non-triviality on the other. We leave this to future research.

Conclusions

In this paper we have presented a new method for recommending research collaboration partners. By aggregating over multiple predictors, the accuracy of our recommendations (measured by comparing recommendations with actually realized collaborations in a later period) is remarkably high. Similar methods could also be used for, for instance, forecasting short term developments in citation networks (Shibata, Kajikawa, & Sakata, 2012).

A surprising result of our study was the relative importance of the predictors, in that an individual predictor's success rate is less clearly related to the predictor's importance than anticipated. The most likely explanation is the fact that the random forest classifier determines relative importance by looking at all predictions, not just the most likely ones. In general, weighted and topology-based predictors are more important than local (neighbour-based) ones.

By focussing on cities and regions this article contributes to the emerging subfield of spatial or regional scientometrics (Frenken, Hardeman & Hoekman, 2009).

References

Adamic, L. & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211-230.

Antonellis, I., Garcia-Molina, H., & Chang, C.C. (2008). Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the 34th International Conference on Very Large Data Bases*, (pp. 408–421). Auckland, New Zealand.

Boshoff, N. (2010). South–South research collaboration of countries in the Southern African Development Community (SADC). *Scientometrics, 84*(2), 481–503.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.

Egghe, L., & Rousseau, R. (2003). A measure for the cohesion of weighted networks. *Journal of the American Society for Information Science and Technology, 54*(3), 193–202.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics. Towards a cumulative research program. *Journal of Informetrics*, *3*(3), 222-232.

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 E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2011 Conference* (pp. 249–260). Durban: ISSI, Leiden University, University of Zululand.

Guns, R. (2012). *Missing links: Predicting interactions based on a multi-relational network structure with applications in informetrics*. Doctoral dissertation, Antwerp University.

Guns, R. & Rousseau, R. (2013). Predicting and recommending potential research collaborations. In J. Gorraiz et al. (Eds.), *Proceedings of ISSI 2013* (pp. 1409–1418). Vienna: AIT.

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* (pp. 538–543). New York: ACM.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.

Langville, A.N., & Meyer, C.D. (2005). A survey of eigenvector methods for web information retrieval. *SIAM Review*, *47*(1), 135–161.

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.

Newman, M.E.J. (2001). Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Platt, J.C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: A.J. Smola et al. (Eds), *Advances in Large Margin Classifiers*, (pp. 61–74), Cambridge: MIT Press.

Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. *Information Processing & Management*, *12*(5), 297–312.

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.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the American Society for Information Science and Technology*, *63*(1), 78–85.

The STIMULATE-6 Group (2007). The Hirsch index applied to topics of interest to developing countries. *First Monday*, *12*(2). http://www.firstmonday.org/issues/issue12_2/stimulate/

Van Eck, N.J., & Waltman, L. (2007). VOS: a new method for visualizing similarities between objects. In H.-J. Lenz, & R. Decker (Eds.), *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (pp. 299-306). Springer.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

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

Yang, LY. & Jin, BH. (2006). A co-occurrence study of international universities and institutes leading to a new instrument for detecting partners for research collaboration. *ISSI Newsletter*, *2*(3), 7–9.