Algorithms for analyzing evolving networks on the Dark Web & in science

Boekhout, H.D.

Part III

Scientific co-authorship networks

5

Evolution of the world stage of global science from a scientific city network perspective

This chapter investigates the stability and evolution of the world stage of global science at the city level by analyzing changes in co-authorship network centrality rankings over time. Driven by the problem that there exists no consensus in the literature on how the spatial unit "city" should be defined, we first propose a new approach to delineate so-called scientific cities. On a high-quality Web of Science dataset of 21.5 million publications over the period 2008–2020, we study changes in centrality rankings of subsequent 3-year time-slices of scientific city co-authorship networks at various levels of impact. We find that, over the years, the world stage of global science has become more stable. Additionally, by means of a comparison with degree respecting rewired networks we reveal how new co-authorships between authors from previously unconnected cities more often connect "close" cities in the network periphery.

This chapter is based on:

- H. D. Boekhout, E. M. Heemskerk, and F. W. Takes. Evolution of the world stage of global science from a scientific city network perspective. In *Complex Networks & Their Applications X*, pages 142–154. Springer International Publishing, 2022

5.1 Introduction

A prevalent way of studying the global science system, is to produce rankings. This may involve rankings of universities based on, for example, publications, scientific impact, collaboration, open access and gender balance in order to "assess university performance on the global stage" [143, 153]. It could involve ranking authors based on, for example: fractionally counted citations [32], the h-index [49], or PageRank in co-citation networks [57]. Or it may involve ranking geographical areas such as countries or cities based on, for example, scientific output [31, 38] or domestic vs. international co-authorship [100]. In short, there are many "levels" at which rankings are produced as part of the study of the science system. In this research, we consider rankings at the city level. However, there exists no consensus in the literature on how the spatial unit "city" should be defined [50]. Therefore, we propose a new approach to delineate *scientific cities*, agglomerations of cities within a small radius based on geo-located addresses on scientific publications.

Related work on the science system often considers measures that are directly computable for a given entity, such as their scientific output, scientific impact, etc. [31, 32, 38, 49, 100, 143]. These directly computable measures usually say little about what position the entity takes within the global science system. Instead, in this research we rank based on the position of the nodes in the *co-authorship network* underlying the science system, in particular using network centrality measures. We do so with the goal of establishing which cities take a "central" role on the world stage of global science, similar to, e.g., related work by Ding et al. [57].

In this chapter, we study changes in centrality rankings in co-authorship networks over time, where our nodes are cities rather than authors and edges denote co-authorship between these cities. By studying the change in centrality rankings we shed light on the stability of the network over time. In particular, we aim to study the stability of the co-authorship networks over time at various levels of "prominence", measured through publication impact in terms of citations received. To this end, we measure the change in network-based rankings that occurs for three co-authorship network variants covering all publications, the top 10%, and the top 1% highly cited publications. Furthermore, we aim to validate the significance of the observed changes in rankings by comparing the changes in these evolving "real-world" networks to changes that occur when the rewiring is performed in a random manner. However, because a sensible rewiring is non-trivial, we propose a new suitable approach to generating rewired networks.

We extract evolving co-authorship networks of scientific cities from a large and high-quality dataset (21.5 million publications with complete author affiliation linkages and geolocation information for the period 2008–2020). With these networks we show that

the world stage of global science has become more stable, and by extension the city co-authorship networks less prone to structural change, over the years. Additionally, we show that city networks follow the expected pattern of more often establishing new co-authorship relations with "close" cities than "distant" cities. Finally, we conclude that, compared to our null model, changes in the network more often occur in the periphery.

In short, we do the following: (1) we propose a new approach to delineating scientific cities; (2) we study changes in various network centrality rankings over time at various levels of "prominence" to study the "world stage of global science"; and (3) we propose a new rewiring null model and determine the significance of the observed changes in centrality rankings by comparing with this null model.

The remainder of this chapter is structured as follows. In Section 5.2 we present our basic network notation and define various network centrality measures and rank correlation measures used in our experiments. Then, in Section 5.3 we discuss how the co-authorship networks were extracted and how we generate randomly rewired networks. Next, the experimental setup, results and limitations are discussed in Section 5.4. Finally, in Section 5.5 we summarize and conclude.

5.2 Definitions, measures, and background

In this section we provide basic network notation and terminology in Section 5.2.1, define centrality measures for ranking vertices in Section 5.2.2 and specify correlation measures to compare those rankings in Section 5.2.3.

5.2.1 Network notation and definitions

In this chapter, we study city co-authorship networks which model scientific cities as nodes and co-authorship on scientific publications by authors from different cities as edges (see Sections 5.3.2–5.3.4 for more details on the specific data used as input). Because the scientific co-authorship relation is an undirected relation, we model networks in this study using an undirected graph $G = (V, E, \omega)$, with V the set of vertices or nodes, E the set of edges $\{u, v\}$ with $u, v \in V$, and weight function ω. We use $n = |V|$ and $m = |E|$. No self-loops and no parallel edges are assumed. For weighted graphs, edge weights are a function of the connected nodes, denoted $\omega(u, v)$, with $\omega(u, v) > 0$ iff $\{u, v\} \in E$ and $\omega(u, v) = 0$ iff $\{u, v\} \notin E$. For unweighted graphs, $\omega(u, v) = 1$ for all $\{u, v\} \in E$. We define a *θ-minimum-weight* graph $G^{\geq\theta}(V, E')$ as the unweighted graph induced from a weighted graph G where $\{u, v\} \in E'$ iff $\omega(u, v) \geq \theta$.

Let $u \rightsquigarrow v$ denote the existence of a *path* between nodes $u, v \in V$. We call $H = (V' \subseteq V, \{\{u, v\} : u, v \in V' \wedge \{u, v\} \in E\})$ a *connected component* when for all

$u, v \in V'$ it holds that $u \rightsquigarrow v$, i.e., all nodes are reachable from every other node. The largest connected component in a graph, in terms of nodes, is referred to as the *giant component*.

The *distance* between two nodes is denoted as $d_G(u, v)$ (with $u, v \in V$) and indicates the length of a *shortest path*, i.e., a path from u to v where the sum of the weights of the edges is minimal. We define the distance between a node and itself as zero, i.e., $d_G(u, u) = 0$. The number of shortest paths connecting $u, v \in V$ is denoted by σ_{uv}, with the number of shortest paths including node $w \in V \setminus \{u, v\}$ denoted by σ_{uwv}.

The *neighborhood* $N_G(v)$ of a node $v \in V$ is defined as the set of nodes to which v links, i.e., $N_G(v) = \{w \in V : \{v, w\} \in E\}$. The *degree* of a node equals the size of its neighborhood, i.e., $deg_G(v) = |N_G(v)|$.

Because we want to study the change in rankings over time, we can accomplish this by considering a series of static *time-slices*, i.e., static networks covering only a few successive years of data. The extraction of these time-slices given our data is discussed in Section 5.3.4.

5.2.2 Ranking measures

In this research we determine the rankings of nodes based on various centrality measures. Specifically, we consider degree, eigenvector, closeness and betweenness centrality. Below we define each of these diverse measures and provide the rationale of high (or low) rankings for cities, with respect to the role these nodes play within the structure of the scientific city co-authorship networks.

Degree centrality

Degree centrality assumes that those nodes with connections to more neighbors are more central. In city co-authorship networks, a high rank translates to co-authorships with many different cities. It is defined as

$$dc_G(u) = \frac{deg(u)}{n-1}. \tag{5.1}$$

Eigenvector centrality

Eigenvector centrality is based on the idea that an actor is more central if it is connected to many actors that are central themselves [125]. Thus, it considers not only the number of adjacent vertices, but also their value of centrality. It can be computed by iteratively setting the eigenvalue ($EV(u)$) of all nodes $u \in V$ to the average of its neighbors, where the initial values of $EV(u)$ are proportional to the degrees of the

nodes, normalizing after each step. We put

$$ec_G(u) = EV(u). \tag{5.2}$$

In scientific city co-authorship networks, a high rank indicates that that city forms co-authorships with cities that co-author with many other cities.

Closeness centrality

Closeness centrality is a measure of how close a vertex is to all other vertices in the graph. As we will be dealing only with the giant component of undirected networks in our experiments, we can employ the simplest version of this measure, as first introduced by Bavelas [13], defined as follows:

$$cc_G(u) = \frac{n-1}{\sum_{v \in V} d_G(u,v)}. \tag{5.3}$$

In other words, the closeness centrality of u is the inverse of the average (shortest-path) distance from u to any other vertex in the graph. In our city networks, highly ranked cities are the cities who require the fewest "intermediary cities" for establishing co-authorships with every other city in the network, i.e., the world.

Betweenness centrality

Betweenness centrality is a measure of the ratio of shortest paths, between any two nodes in the network, that a node lies on [33]. In other words, it measures the extent to which shortest paths pass through a specific node. It is defined as follows:

$$bc_G(u) = \sum_{u \in V \setminus \{s,t\}} \frac{\sigma_{sut}}{\sigma_{st}}. \tag{5.4}$$

In city co-authorship networks, lying on a shortest path connecting two cities indicates that establishing a co-authorship between those cities may most easily be accomplished through an introduction or collaboration with your city. Thus, highly ranked cities may form an important factor in brokering new co-authorships between "distant" cities.

5.2.3 Rank comparison measures

In order to systematically compare two rankings of nodes in evolving networks, a measure is required that can express their (in)equality in one normalized number. Two correlation-based measures suited to this task are the *Spearman* and *Kendall's Tau* rank correlations. The advantage of applying the Spearman rank correlation is that the exact difference in ranking between all pairs of nodes in both rankings is taken into account [89, 131]. On the contrary, Kendall's Tau correlation considers only the extent to which the pairs of nodes are identically ordered.

5.3 Materials and methods

In this section we first discuss the bibliographic database from which we extract our co-authorship networks in Section 5.3.1. Then, Section 5.3.2 describes our new approach for the delineation of the scientific city agglomerations, i.e., our nodes. Section 5.3.3 discusses the publication sets and counting methods used to compute the edge weights. Next, we describe how we obtain our final co-authorship network time-slices in Section 5.3.4. Finally, in Section 5.3.5 we explain how we generate the random networks.

5.3.1 Bibliographic database

Our analysis is based on Clarivate's Web of Science database (WoS). Specifically, we use the in-house version of WoS at the Centre for Science and Technology Studies (CWTS) at Leiden University from April 2021. This version of WoS has been enriched with its own: citation matching; assignment of publications to universities and organizations in a consistent and accurate manner [143]; geocoding of the author addresses; and improved author disambiguation [37].

We consider publications published in 2008–2020 categorized as Article, Review, Letter or Proceeding Paper. Publications with missing author-affiliation linkages or missing both geolocation and organization information are excluded. This leaves 21.5 million publications (87.2% of total), covering 196 countries.

5.3.2 Delineation of scientific city agglomerations - nodes

Csomós [50] discusses the various challenges of spatial scientometrics focusing on the city level. One of these challenges is that there exists no consensus in the literature on how the spatial unit "city" should be defined and how metropolitan areas should be delineated. Alternative spatial units that lie between the level of the organization and the country are sublevel territorial units, such as federal states or Metropolitan Statistical Areas (USA), provinces (Canada), and NUTS regions (EU member states). Although these sublevel territorial units provide fair/consistent spatial units for studies focusing on specific countries or regions, their definitions can differ drastically and are therefore not suitable for global studies such as the one conducted in this chapter. Csomós posited that studies that focused on creating urban agglomerations are considered to be the most trustworthy. In one such study, Maisonobe et al. [100] asserted that the aim of urban agglomerations was to produce universal criteria instead of national criteria. As this makes urban agglomerations suitable for global studies, we follow this line of reasoning for our study.

Our approach to constructing a set of urban agglomerations (cities) most closely matches that described by Maisonobe et al. [101]. However, whereas their approach agglomerates to metropolitan areas the size of world cities, we instead agglomerate to smaller clusters of research localities, which we call *scientific cities*. Here, we rely on the observations of Bornmann and de Moya-Anegón [31] and Catini et al. [38] that “institutions are frequently spatially clustered in larger cities" and that “research institutions involved in scientific and technological production” are generally located close to each other and produce well-outlined research clusters within cities. By segregating world cities with multiple research clusters, instead of merging smaller urban areas to the level of world cities, we allow for insights with potential relevance for policy making at the level of research clusters. Furthermore, one goal of studies at the city level is to consider local collaboration outside the organization/institution, where being in the same city would indicate easier collaborations due to short travel times. However, travel times between urban agglomerations at the level of world cities are highly dependent on the available infrastructure. By focusing on smaller urban agglomerations, we ensure relatively consistent travel time within each scientific city.

Following these observations, we segregate world cities with multiple research clusters so that, as long as research clusters are spatially further than eight kilometers apart, each research cluster is considered its own scientific city. This approach allows us to create globally comparable geographical entities, similar to [101], whilst still delineating between distinct cities in notoriously difficult regions for agglomeration such as “de Randstad" in the Netherlands (Amsterdam, Leiden, The Hague, Delft, etc.). A manual inspection of a sample of regions found that this eight-kilometer radius works well throughout most regions of the world, often dividing world cities into their named districts. One clear exception, that can be considered a limitation of this approach, is that Chinese cities, such as Beijing, do not appear to allow for segregation into research clusters as the addresses listed on such publications tend to be at the municipality level, which in the case of Beijing covers approximately 16,000 km^2 [50]. Consequently, these cities may come to have an advantage with respect to their scientific output over other world cities for which we are able to delineate multiple scientific cities. To reduce the number of cities with very low scientific output in our set of scientific agglomerations, we merged cities with fewer than ten publications into the closest scientific agglomeration with more than a hundred publications (within 30 kilometers).

Note that in order to generate the set of scientific cities and subsequently assign each author for each publication to one of them, we have access to two forms of geolocation information. First, we have author address geolocation information; and second, we have organization geolocation information through the author-affiliation linkages. However, we do not always have access to both for all authors/publications. It seems reasonable to assume that the author address geolocation is the most likely to

correspond to the actual location of their workplace. On the contrary, organizations may be spread over multiple buildings and locations, which can lead to inaccuracy. Therefore, we initially generate the set of scientific cities based solely on the available author address information. Subsequently, we match organizations to scientific cities based on their co-occurrences for specific authors on publications. For this process, we consider both the frequency of the co-occurrences and the distance between the organization geolocation and the scientific city, i.e., geolocation based on author address. When there exists a scientific city for which an organization has both the highest frequency and smallest distance, then it is assigned to that scientific city. If no such case exists, then the scientific city with the highest co-occurrence frequency is chosen. The few remaining organizations, which had two (or more) highest co-occurrence frequencies, are then resolved manually, which in some cases involves the creation of new scientific cities. In the end we are left with 16,619 distinct scientific cities, with co-authorships forming 2,084,123 distinct city pairs. These distinct cities and city pairs form respectively the (potential) nodes and edges of our networks.

5.3.3 Publication sets and counting methods - edge weights

Recall that we aim to understand differences between all scientific publishing activity and high impact publishing activity. Therefore, edges and their weights are determined based on three different publication sets:

1. all, the full publication set;
2. hcp_{10}, consisting only of the top 10% highly cited publications; and
3. hcp_1, consisting only of the top 1% highly cited publications.

The highly cited publication sets are determined by ranking publications within each respective publication year and Web of Science (WoS) subject category [47]. The publications are ranked by the number of citations they received in the first three years after publication excluding self-citations, where ties are broken by the number of self-citations. By ranking separately for publication years we are able to determine edge weights for each respective year, thus allowing the creation of network time-slices. Additionally, we rank each WoS subject category separately, because different scientific fields have different citation practices and the WoS subject categories can be used as a proxy for scientific fields. If we were to rank irrespective of field, we would erroneously select too many publications from certain fields that receive on average more citations per publication than in other fields, such as Biochemistry & Molecular Biology [145].

Deciding on the right counting method for a given purpose was another challenge highlighted by Csomós [50]. For example, when comparing the scientific output of

Table 5.1: Basic network statistics (see Section 5.3.4 and 5.4.1 for details).

		all			hcp_{10}			hcp_1		
time-slice	θ	n	m	avg deg	n	m	avg deg	n	m	avg deg
2008–2010	3.56	5,870	71,051	24.2	1,719	10,991	12.8	324	749	4.6
2009–2011	3.75	5,920	72,346	24.4	1,740	11,023	12.7	321	748	4.7
2010–2012	3.96	5,927	72,919	24.6	1,767	10,971	12.4	316	743	4.7
2011–2013	4.27	5,820	71,995	24.7	1,727	10,764	12.5	311	735	4.7
2012–2014	4.60	5,822	72,194	24.8	1,728	10,670	12.3	303	697	4.6
2013–2015	4.97	5,845	72,696	24.9	1,741	10,771	12.4	313	726	4.6
2014–2016	5.37	5,811	71,774	24.7	1,678	10,492	12.5	301	704	4.7
2015–2017	5.72	5,708	71,240	25.0	1,707	10,305	12.1	287	694	4.8
2016–2018	6.01	5,709	71,229	25.0	n/a	n/a	n/a	n/a	n/a	n/a
2017–2019	6.18	5,700	72,049	25.3	n/a	n/a	n/a	n/a	n/a	n/a
2018–2020	6.32	5,723	73,433	25.6	n/a	n/a	n/a	n/a	n/a	n/a

cities it may be desirable to count a publication that involves multiple cities fractionally towards each city. Traditionally fractional counting assigns equal size parts of the publication to each city. Here, we use the completeness of the author-affiliation linkages in our data to perform fractional counting based on the number of authors linked to each scientific city. Hereby, we aim to assign fractions representing the expected contribution of each city or city-pair.

In this study we use *city-pair fractional counting* for determining the edge weights. Let $na_{i,j}$ indicate the number of authors on a publication i linked to scientific city j and let C be the set of contributing cities. The fraction of publication i assigned to city pair $a, b \in C$ is then determined by $(na_{i,a} \cdot na_{i,b})/(\sum_{j,k \in C} na_{i,j} \cdot na_{i,k})$.

5.3.4 Network formation

We are now ready to extract the various co-authorship network time-slices. Due to variations in the time between conducting research and the publication of that research, there exist small annual fluctuations in scientific activity. A common approach to account for these fluctuations is to compute a normalized or moving average over a span of three years [100]. Therefore, the 11 time-slices we extracted each covers three years, respectively 2008–2010, 2009–2011, . . ., 2018–2020.

For each time-slice and publication set, edge weights are determined using city-pair fractional counting (see Section 5.3.3). Next, we retain only those edges with a summed weight of more than one per million total publications in that time-slice, i.e., we obtain the θ-minimum-weight graphs with $\theta = \#publications/1,000,000$. Thus, we exclude edges representing city collaborations that we deem too weak, while accounting for the overall increase in the number of publications per year in general. Finally, we exclude nodes and edges from the networks that are outside their giant components. Some basic statistics of the resulting networks are given in Table 5.1.

5.3.5 Evolving degree respecting rewired networks

In order to better understand the changes observed in the centrality rankings between subsequent time-slices, we want to compare these changes to those observed if the network rewiring was done randomly. The procedure of generating these networks (see Algorithm 9) involves rewiring a previous time-slice with an equal number of edge removals (lines 2–10) and edge additions (lines 11–27) as performed in the evolution of the real-world network to the subsequent time-slice. During this procedure we aim to retain the degree distribution of the real-world time-slices as close as possible (lines 5 and 15–21). Therefore, a comparison with this null model highlights where in the city co-authorship networks (core, periphery, etc.) many real-world structural changes occur. We call these networks *evolving degree respecting rewired* (EDRR) networks. Note that robustness checks for the constants used in Algorithm 9 are our of scope of this research and may be performed in future work.

5.4 Results

In this section we discuss our experimental setup, results and limitations.

5.4.1 Experimental setup

For each centrality measure (see Section 5.2.2) and impact level, we include in the rankings only those cities that occur in all time-slices. Additionally, for the full publication set (all) we consider for each time-slice only the top 2000 cities. This ensures that cities that do not play a "central" role in the networks, i.e., that are not of direct importance to the world stage of global science, are excluded. Thus, it mirrors the natural filtering that occurs for the hcp publication sets. By correlating centrality rankings of the more central cities, observed changes and differences are more relevant for understanding the world stage of global science.

Because we use three years of citations after publication for determining the hcp publication sets, the last three time-slices (2016–2018, 2017–2019 and 2018–2020) are excluded from the analysis for hcp_{10} and hcp_1. For this same reason, statistics on these time-slices are excluded from Table 5.1.

5.4.2 Centrality changes over time at various levels of impact

Figure 5.1 shows the correlations between subsequent time-slices for each publication set and for the four centrality measures under consideration. For all four measures we see that the correlations for all are slowly but steadily rising. This tells us that over

Algorithm 9 Algorithm for generating an EDRR network

Require: Previous time-slice $G_p = (V_p, E_p)$ and current time-slice $G_c = (V_c, E_c)$
Ensure: EDRR network G_r

$G_r \leftarrow G_p$
for $\{u, v\} \in E_p$ **and** $\{u, v\} \notin E_c$ **do**
  $E_{poss} \leftarrow \emptyset$, $r \leftarrow 0.1$
  while $E_{poss} = \emptyset$ **do**
    $E_{poss} \leftarrow \{\{s, t\} \mid \{s, t\} \in G_r$ and
      $(1 - r) \cdot deg_{G_p}(u) \leq deg_{G_r}(s) \leq (1 + r) \cdot deg_{G_p}(u)$ and
      $(1 - r) \cdot deg_{G_p}(v) \leq deg_{G_r}(t) \leq (1 + r) \cdot deg_{G_p}(v)\}$
    $r \leftarrow r \cdot 2$
  end while
  $e_r \leftarrow$ random element from E_{poss}
  $G_r \leftarrow G_r \setminus e_r$
end for

$n\_new \leftarrow [u^{deg_{G_c}(u)}$ for all $u \in V_c, u \notin V_p]$
for $\{u, v\} \in E_c$ **and** $\{u, v\} \notin E_p$ **do**
  $E_{poss} \leftarrow \emptyset$, $r \leftarrow 0.1$
  while $E_{poss} = \emptyset$ **do**
    if $deg_{G_p}(u) = 0$ **then**
      $E_{poss} \leftarrow \{\{s, t\} \mid \{s, t\} \notin G_r$ and
        $s \in n\_new$ and
        $(1 - r) \cdot deg_{G_p}(v) \leq deg_{G_r}(t) \leq (1 + r) \cdot deg_{G_p}(v)\}$
    else if $deg_{G_p}(v) = 0$ **then**
      $E_{poss} \leftarrow \{\{s, t\} \mid \{s, t\} \notin G_r$ and
        $t \in n\_new$ and
        $(1 - r) \cdot deg_{G_p}(u) \leq deg_{G_r}(s) \leq (1 + r) \cdot deg_{G_p}(u)\}$
    else
      $E_{poss} \leftarrow \{\{s, t\} \mid \{s, t\} \notin G_r$ and
        $(1 - r) \cdot deg_{G_p}(u) \leq deg_{G_r}(s) \leq (1 + r) \cdot deg_{G_p}(u)$ and
        $(1 - r) \cdot deg_{G_p}(v) \leq deg_{G_r}(t) \leq (1 + r) \cdot deg_{G_p}(v)\}$
    end if
    $r \leftarrow r \cdot 2$
  end while
  $e_r = (u_r, v_r) \leftarrow$ random element from E_{poss}
  $G_r \leftarrow G_r \cup e_r$
  $n\_new \leftarrow n\_new \setminus \{u_r, v_r\}$
end for

the years the full world stage of global science has become increasingly more stable, suggesting the city co-authorship network has become less prone to structural change. Most "stabilisation" appears to have occurred between 2009 and 2015 (time-slices 2008–2010 and 2014–2016) and is most pronounced for betweenness centrality. Thus, annual changes in the city co-authorship network have had an increasingly diminished effect on shortest paths in the network. A pessimistic interpretation of this observation

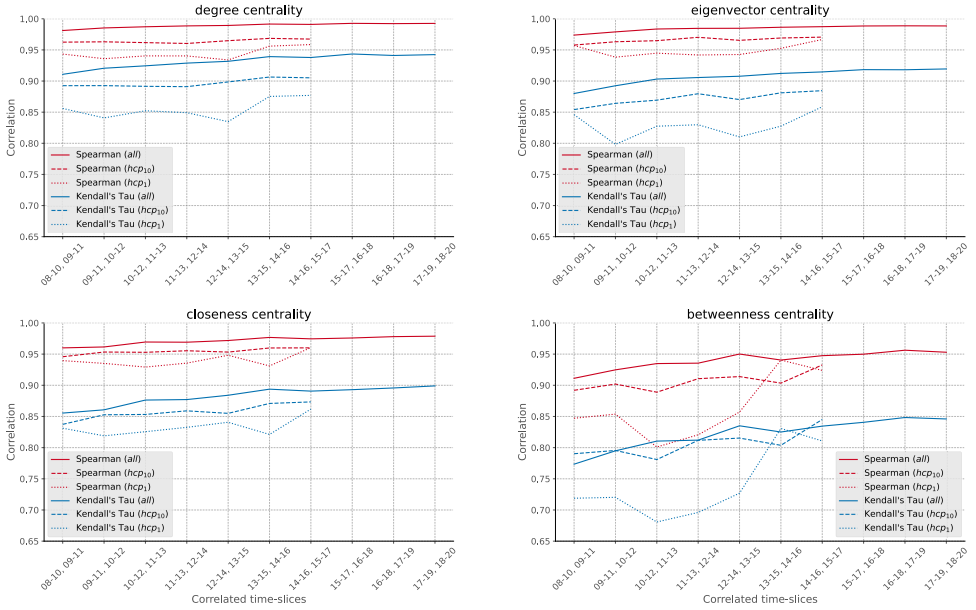

Figure 5.1: Rank correlations of subsequent time-slices at varying levels of impact.

may be that fewer "meaningful" bridging collaborations appear to be formed between "distant" (clusters of) cities.

In Figure 5.1 we observe lower correlations for publication sets representing relatively higher impact. A (partial) explanation for this can be found in the nature of the construction of the hcp networks as well as in their respective size and average degree (see Table 5.1). Because the same θ value is used for each publication set, it is significantly harder for a city co-authorship relation for hcp_1 to be considered "meaningful", than for all, since there are a hundred times fewer total publications. As a result, it is to be expected that the "core" of the city co-authorship network is more significantly affected on a yearly basis, which in turn affects the rankings. Furthermore, because the average degree for hcp_1 networks is quite low, relatively weak co-authorships that connect "distant" (clusters of) cities are more likely to (dis)appear from the hcp networks without alternative short paths connecting them, thereby significantly impacting the rankings for centrality measures such as closeness and betweenness. Indeed, we see especially large fluctuations in the trend of correlations for betweenness centrality for hcp_1.

5.4.3 Real-world rank correlation vs. null model

When comparing with randomly generated networks, sufficiently many random networks are required to establish meaningful differences between the random and real-world networks. Therefore, we generated 100 EDRR networks, using the process described in Section 5.3.5, for each publication set and time-slice (except the first). For each EDRR network we computed the correlation between the rankings for that EDRR network and the real-world network of the previous time-slice. Figure 5.2 shows the real-world correlation alongside the mean and the error range defined by the standard deviation (sd-range) of the correlations for each set of EDRR networks. Because we expect confounding effects from our EDRR network generation procedure and it is a local measure, degree centrality is excluded.

For eigenvector centrality we observe that the real-world correlations often lie within the sd-range of the EDRR correlations for both Spearman and Kendall's Tau correlations. Although Kendall's Tau correlation for all is almost consistently above random, the difference can hardly be called significant.

For closeness centrality we see that all publication sets have Kendall's Tau correlations that are almost consistently above random, while Spearman correlations are around, above and below random at times. However, the trends of Spearman and Kendall's Tau correlations have similar shapes. This implies that while the real-world networks observe fewer changes in the order of pairs of nodes than the EDRR networks, this difference is negated by the exact difference in the rankings. In other words, the real-world city networks observe many but relatively small changes in rank while the EDRR networks observe more substantial changes in rank, i.e., the EDRR networks more often remove/add edges connecting otherwise "distant" clusters of cities while the real-world networks remove/add edges between cities that are otherwise already considered "close". In short, the real-world city networks follow the expected pattern of more often establishing new co-authorship relations with "close" cities than "distant".

For betweenness centrality we observe a similar trend as for closeness centrality. Especially for hcp_{10} the difference between the real-world Kendall's Tau correlation and random is far more significant than it was for closeness. In other words, changes in the real-world networks appear to have far less influence on the betweenness centrality than in the EDRR networks. This may imply that more of the annual real-world city network rewiring occurs in the periphery. This inference is further supported by the fact that the differences are smaller for the full publication set for which most periphery nodes are likely already excluded from the analysis. Thus, the real-world removal and addition of edges impacts the shortest paths between all pairs of cities less than random.

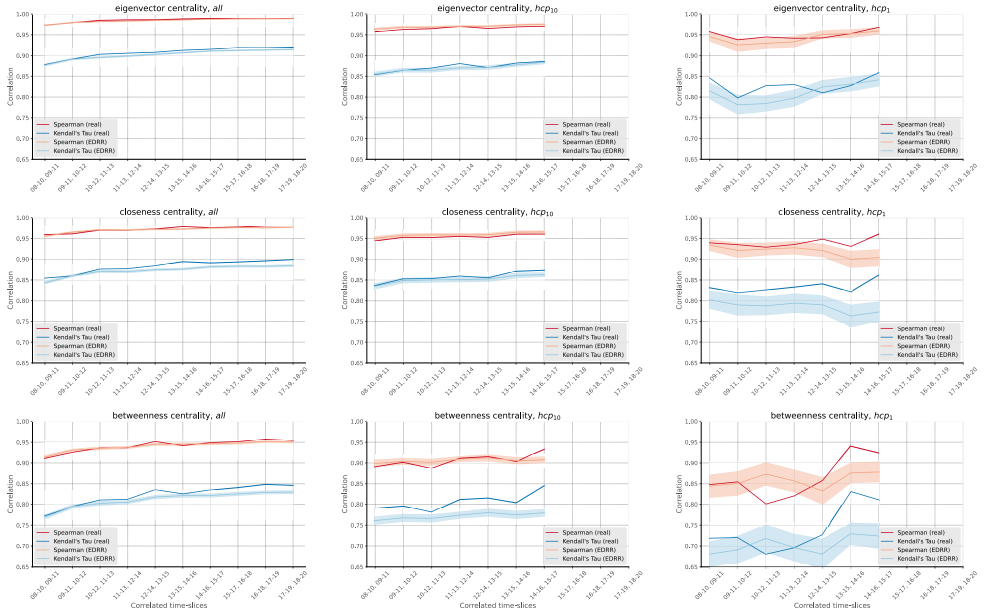

Figure 5.2: Rank correlations of subsequent time-slices comparing the real-world network results against the mean and sd-range of EDRR networks.

5.4.4 Limitations

An important limitation of this research is that we focus entirely on unweighted networks, generated from weighted networks using thresholds. When studying the co-authorship between cities, the ideal edge weight should represent the ease with which a co-authorship between cities is formed. Although we have a fractional publication output associated with each edge, this weight is likely a poor representation of the ease of co-authorship formation. Since there is no simple numerical computation of "the ease of forming a co-authorship" based on existing scientometric data, we instead use θ-minimum-weight graphs to establish a minimum co-authored scientific output for a relation to be considered "meaningful".

5.5 Conclusions

In this chapter we investigate the stability and evolution of the world stage of global science at the city level by analyzing changes in network centrality rankings over time. First, we proposed an approach for delineating scientific cities and extracted 3-year time-slices of scientific city co-authorship networks from Web of Science at various

levels of impact. Comparing correlations between centrality rankings of subsequent time-slices, we determined that the world stage of global science has become more stable over time. We propose a new rewiring procedure to generate so-called EDRR networks in order to determine significant real-world rank correlations compared to a sensible null model. We found that closeness and betweenness centrality rankings were more stable for the real-world networks, implying that new co-authorships between authors from previously unconnected cities more often connect "close" cities in the network periphery. Having established a systematic method of comparing centrality rankings over time, we want to find more substantive insights for specific cities in future work.

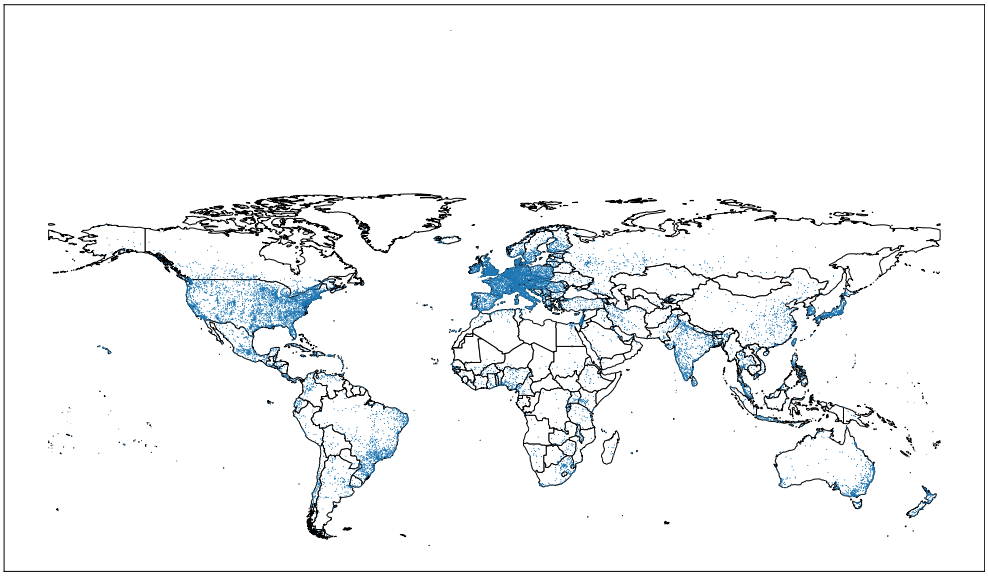

Figure 5.3: All scientific cities plotted on a global map.