EPJ manuscript No.
(will be inserted by the editor)

Collaboration range: effects of geographical proximity on article impact

Andrea Apolloni[1,a], Jean-Baptiste Rouquier[1,2], and Pablo Jensen[1,b]

[1] Institut des Systèmes Complexes Rhône-Alpes (IXXI) and Laboratoire de Physique, UMR 5672, École Normale Supérieure de Lyon, 69007 Lyon, France
[2] Université de Lyon, École Normale Supérieure de Lyon, LIP, UMR 5668, France

Abstract Spatial scientometrics studies how geography influences knowledge creation. In the recent years there has been a surge in this kind of studies, due to the increase of international collaborations. Most of the work in this field has been focused on the geographical distribution of researchers, whilst few have considered how proximity between coauthors influences research quality. In this work we leverage a dataset of geolocalized articles to assess the effect of geographical distance on article impact.
More precisely, the dataset, provided by the Observatory of Science and Technology (O.S.T.), consists of roughly 10^6 scientific articles, gathering *all* European articles written in 2000 and 2007, spanning 9 disciplines. We evaluate under which geographical extent coauthorships have higher probability of resulting in high impact articles ("high impact" is here approximated by "being in the top 10% most cited articles of its discipline"). We also describe spatial distribution of coauthorship, delineating geographical areas where the production is proportionally higher. The distribution is evaluated both in term of km (as the crow flies), and in terms of administrative partitions (authors' cities, regions, countries).

1 Introduction

Spatial scientometrics aims to study where and under which conditions knowledge is created and transferred. Contrary to infectious diseases and economical exchanges, the fluxes of ideas and knowledge are difficult to define and quantify. Most studies have focused on the spatial distribution of patents and articles as a proxy to knowledge creation, and on quantifying the number and types of collaborations. Previous works have shown that knowledge production is geographically localized, but that these areas have a collaborative network spanning over different countries [12]. Moreover, in a time where the tendencies to internationalization and globalization are increasing, the study of how different geographical units are collaborating is gaining momentum both among researchers and among policy makers.

[a] e-mail: `abu.apolloni@gmail.com`
[b] e-mail: `pablo.jensen@ens-lyon.fr`

The work on geographical proximity and publication output has been focused on international distribution and research output [8]. The authors of [10, 8] have noticed that an increase in distance significantly decreases the frequency of research collaborations. However, while there is an obvious bias towards intra-national collaborations, adding an author from a foreign institution adds more citations on average than adding an author from a domestic organization [7]. There are several reasons why distance is playing an important role in coauthorship: first, the needs for face-to-face interactions becomes more expensive as the physical distance increases; second, language, funding, intellectual property rights are country dependant and constrain interaction between institutions; third, social ties between coauthors, like in any social network, are geographically biased [11, 12]. Up to now, few studies has focused on spatial analysis including regional level [7] and while they do, they are limited to some countries, focusing more collaboration frequency than on research quality.

The analysis of coauthorship data has shown that, until 2000, most of the coauthorship were nationally biased [9]. Recently, the European Union (EU) has been fostering the collaboration among scientists of different countries to create a European Research Area (ERA) [7]. To this end, the EU has funded projects involving different countries and eased possible boundaries that hindered collaborations. In the last decade, this has resulted in an increase of international collaborations, reflected in an increase of international coauthorship. However, the link between research quality and geographical proximity is still an open question. One recurrent problem in spatial scientometrics is to define the relevant geographical units, since there is no general consensus that administrative boundaries coincides with relevant ones [8]. This kind of study thus requires a multilevel framework that quantifies the effect of different geographical aggregations.

The main scope of this article is to present a description of the evolution, quantitatively and qualitatively, of the European coauthorship network. We have analyzed a dataset of articles published in 2000 and 2007 with at least one author with a European affiliation. The coauthorship network has been analyzed at two levels (international level and NUTS3 level) in order to show the heterogeneity between both levels. We used the notation of Matthiessen [3] to classify nodes and links of the coauthorship network. We also present some preliminary results about the quality of a link as a function of the euclidean distance.

The article is organized as follows: in Section 2, we detail the characteristic of the dataset used and derive some statistics; in Section 3, we present a coauthorship network at two different levels (international and NUTS3), we also give some insights on the effect of euclidean distance on the visibility of an article; we finally conclude in Section 4.

2 Data set description

The dataset used in this paper has been kindly provided by the OST (*Observatoire des sciences et des techniques*, Sciences and technologies observatory), a French organism compiling bibliographical data and producing comprehensive reports from it. The dataset consists of all articles (referenced by the OST) written in 2000 or 2007, in 9 scientific disciplines (Biology, Medicine, Mathematics, Physics, Geology, Agriculture, Material Science, Engineering, Multidisciplinary), where at least one author has a European affiliation. The total amount of articles is around 10^6, with 4.4 10^5 of them written in 2000 and 6.2 10^5 written in 2007. Figure 1 gives a first glimpse on the dataset, and shows that there has been an increase in each discipline, in particular in Medicine and Biology, which are also the disciplines producing the most articles. The OST parses author affiliations and produces in particular clean data about the

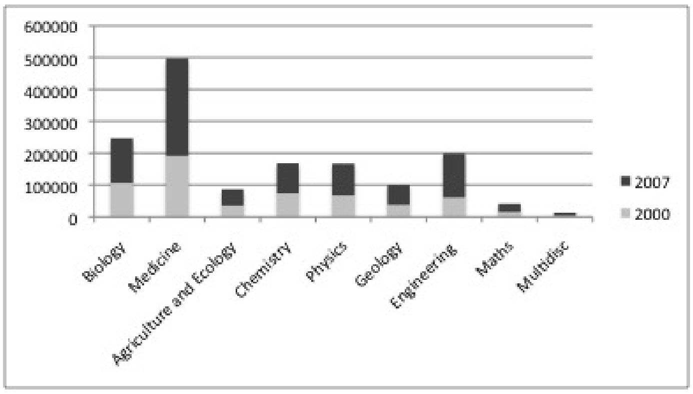


Figure 1. Articles' distribution per year and per discipline

city and administrative regions associated to an article. The administrative regions are defined in terms of NUTS3[1] and are missing for Russia and Switzerland. We then enriched this dataset with the geographical coordinates of each city.

A classical indicator in scientometrics is the number of citations. It has some shortcomings: external factors like friendship relations and Matthew effect (citing articles already cited) [14] prevent the number of citations from being a reliable estimator of the quality of a paper [**?**]. Moreover, the number of citations strongly depends on discipline: different disciplines have very contrasted habits regarding number of publications, and number of references at the end of an article. In order to overcome these difficulties, and to put all disciplines at the same level, we use as estimator of the *visibility* the property of being part of the 10% most cited articles (globally), computed separately for each discipline. For brevity, we simply refer to "Top10" (we also define "Frac10" as the percentage of articles in Top10). The average percentage of articles in Top10 is 7.4% in 2000 and 8% in 2007, which is close to 10% as expected.

Figure 2 shows, for each country and year, the number of articles and Top10 articles having at least one author affiliated in this country. United Kingdom is producing the most articles (20%), followed by Germany (19%). Switzerland has the highest proportion of Top10 (14% in 2000, 15% in 2007). Most importantly, each country has increased both its production and its quality between both years.

To each article is associated the list of authors, cities, institutions, countries and NUTS3 participating to the article. Note that some authors have several affiliations, thus there can be more cities than authors for a given article. Figure 3 gives an overview of the number of distinct cities, NUTS3 or countries per article. Most articles are written in a single city. International articles provide only 10% of the total production in 2000 and 15% in 2007, but 15 % of the Top10 production in 2000 and 20% in 2007. International collaborations have increased 19% in the total production and 25% in the Top10.

To conclude this section, we have looked for the factors influencing the visibility of an article. In general, increasing the number of coauthors, institutions or countries increases the visibility. But this is not systematic: for articles written by authors in the same country, the geographical diversity (i.e. different NUTS3 or cities) has a negative or un-influential effect. More precisely, table 1 shows the results of a multilinear

[1] The *Nomenclature of Territorial Units for Statistics* (NUTS) is a uniform breakdown of spatial units in the EU which follows a four-level hierarchy: NUTS0 (coarse) to NUTS3 (fine). NUTS3 corresponds to a labour pool in most countries. See `http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction/` for details.

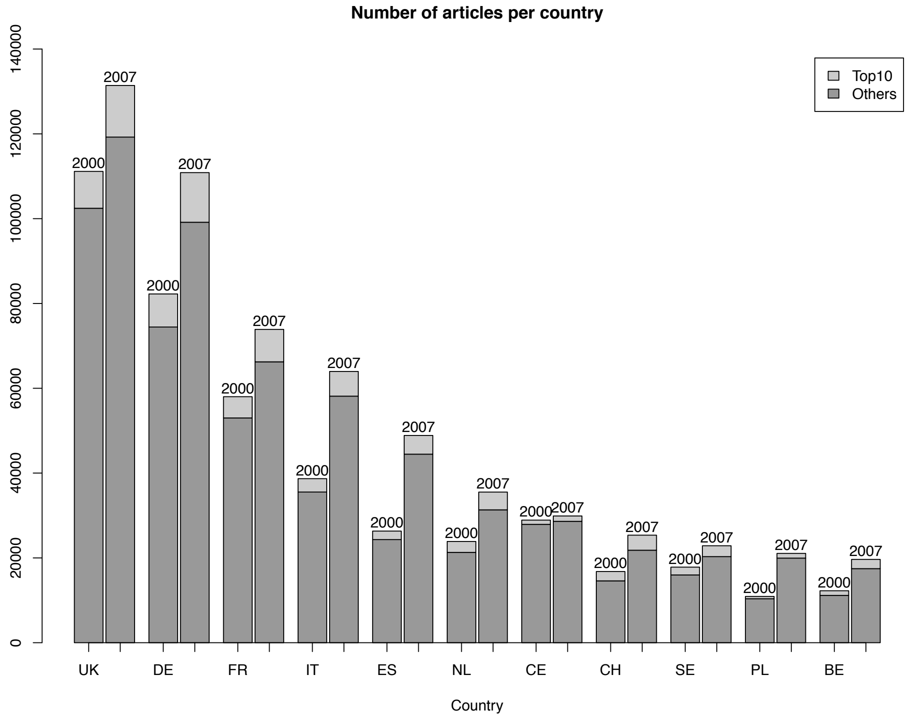


Figure 2. Articles' distribution per year and country. Only countries with more than 10^4 articles are shown.

Table 1. Coefficients for the multilinear regression $\frac{\text{Top10}}{N} \simeq a\ \#\text{Authors} + b\ \#\text{Countries} + c\ \#\text{Affiliations} + d\ \#\text{NUTS3}$ for both 2000 and 2007. The error is evaluated as one standard deviation.

Year	2000	2007
Authors	0.68 ±0.05	0.30 ± 0.04
Countries	1.7 ±0.1	2.7 ± 0.3
Affiliations	1.6 ±0.1	1.6 ± 0.1
NUTS3	-0.3 ± 0.2	-0.3 ± 0.2

regression of Frac10 vs. the number of countries, authors, affiliations and NUTS3. It means that, all other things being equal, adding an author increases the probability of being in the Top10 by 0.68% in 2000 and 0.30% in 2007. Also, when the number of authors is kept constant, the visibility of international articles is much higher, confirming the results of [7].

3 Network analysis

In order to study the effect of collaborations on article's visibility, we have considered a geographical coauthorship network. As in previous works on coauthorship [1, 2], the network is bipartite:

one type of nodes represents the article, the other type a geographical unit (Country, City, NUTS3). A link between both sides is established if the article has been

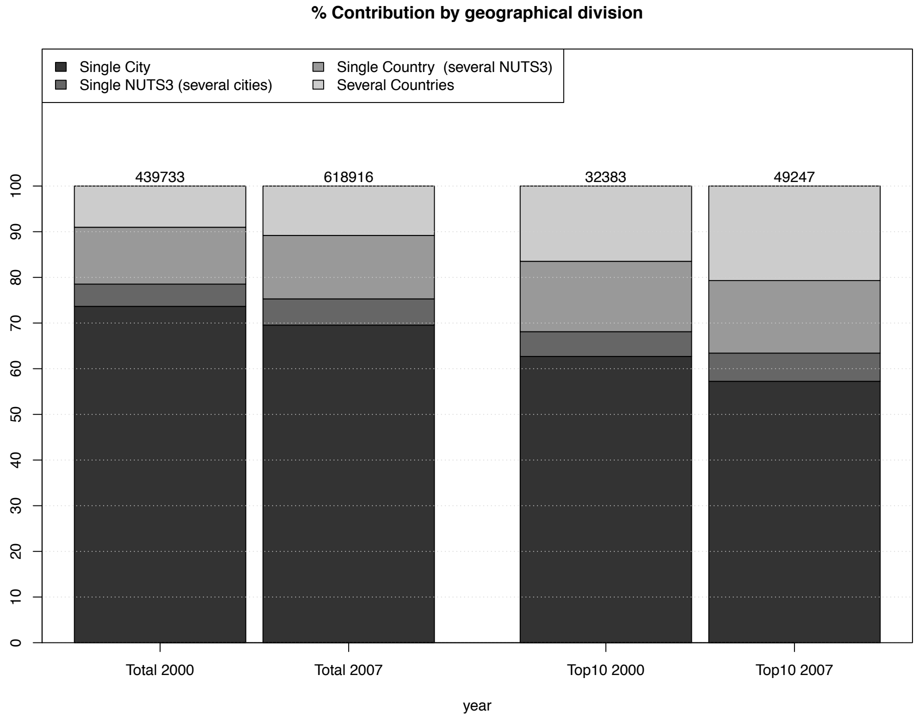


Figure 3. Distribution of articles by geographical divisions.

written by one or more authors in one of the geographical units. A link between two geographical units is established if authors of two different units collaborate in writing an article. It follows that a coauthorship involving more than two NUTS3 induces a cllique on the geographical side. Each link is weighted by the total number of articles involving its endpoints. The *quality* of a link is defined by the fraction of Top10 articles involving its endpoints. The resulting network is then a weighted one and we aim to study the effect of the distance on both the weight and the quality of the links.

This kind of network is a multi-scale model, since nodes can represent different geographical subdivisions (country, regional, city level) and heterogeneities at one level can disappear when aggregating to a higher level. At country level, for example, the network is composed by 27 nodes (the EU27 zone) and 330 links. It represents 1.86 10^5 collaborations in 2000 and 3.41 10^5 in 2007, and the distribution of links is almost uniform.

In order to study possible heterogeneities in this coauthorship network, we chose to analyze the microscopic network at NUTS3 level. As seen in Figure 3, only a small percentage of articles are collaborations of several cities in the same NUTS3, thus aggregating cities at NUTS3 level is almost equivalent to considering the city level. Furthermore, the same city can appear with different names. As pointed out by Bornmann [6], the NUTS3 agregation gets rid of the problem of defining a urban area.

The NUTS3 network consists of 1136 nodes with almost 4.0 10^4 links in 2000, increasing to 5.8 10^4 in 2007. Between 2000 and 2007, NUTS3 have enormously increased the number of collaboration from $\langle k \rangle = 29$ in 2000 to $\langle k \rangle = 100$, and the degree distribution has become more right skewed, thus indicating that more authors from different NUTS3 collaborate for each article. The networks of both years have a low density (0.06 in 2000 and 0.09 in 2007). For both years, the networks are formed by a unique giant component plus a few isolated nodes. If one adds the fact that the clustering coefficient is high in both cases, this supports the idea that coauthorship (meant as the number of copublishing authors from different NUTS3) has enlarged from one year to the other.

Enlarging the size of the coauthorship mean increasing the range at which two NUTS3 collaborate. There are different ways of evaluating distances between geographical units: some of the possibilities are in particular distance as the crow flies, regional distance, national and international collaborations, institutional distance and traveling time [8, 7]. In this study we have considered the euclidean distance, for reason of data availability, and distinguished between national and international links.

Since we are dealing with a longitudinal analysis we classify nodes and links based on variation of properties. We use the notation of Matthiessen & al. [3] for classifying the different nodes and links. Accordingly a node will be of type 1 (*hotspot*) if $\Delta\text{Top10} > 0$ and $\Delta\text{Article} > 0$, of type 2 (*Focus on success*) if $\Delta\text{Top10} > 0$ but $\Delta\text{Article} < 0$, type 3 (*Black Hole*) if $\Delta\text{Top} < 0$ if $\Delta\text{Article} < 0$ and finally type 4 (*reputationloss*) if $\Delta\text{Top10} < 0$ but $\Delta\text{Article} > 0$. We notice that this kind of classification provide informations about the tendency in time of a certain NUTS3, but in some cases can lead to misleading conclusions about the geographical unit visibility (i.e. the Frac10 evaluated over geographical unit production). When necessary, we will use another estimator, the *visibility variation* v:

$$v := 100\,\frac{\text{Frac10(2007)} - \text{Frac10(2000)}}{\text{Frac10(2000)}} \tag{1}$$

A similar classification is used for links between NUTS3: a type 1 link (*hot link*) is a link along which both the weight and the quality have increased; through a type 2 (*success*) link the number of top10 collaborations has increased while the total collaborations have decreased; through a type 3 (*losing*) link both quality and quantity have decreased; through a type4 (*reputationloss*) link only the quality has decreased while the number of collaboration has increased. We name the network composed only link of type 1 and 2 the *excellency network*, and name *neutral* the network with links of all types.

3.0.1 Country level

As we have seen before, all countries have shown an increase in the number of articles and quality, so all of them can be classified of type 1. Although all the countries are hotspots, their visibility has not changed uniformly. For countries like Poland and Romania, the visibility variation v (1) is of order of -10% and -40%, while for Hungary and Turkey it is of order 50%. However these countries have few publications and a low Frac10 in 2000. For countries with a large visibility in 2000, like United Kingdom, Italy, France, Switzerland and Germany, the visibility variation ranges from 6% (Switzerland) to 18% (UK). The visibility of this second set of countries is higher than the European average in 2007 (8.4 %), ranging from 9% (Italy) to 14% (Switzerland).

The excellency network of European countries is shown in Figure 4, the majority of links (80%) are hotlinks. The excellency network is denser and more clustered than the neutral one. From a geographical point of view, the excellency network shows no dependency on distance, although the largest fraction of collaborations are between the UK, France, Italy, Spain and Germany: those countries form a clique and produced

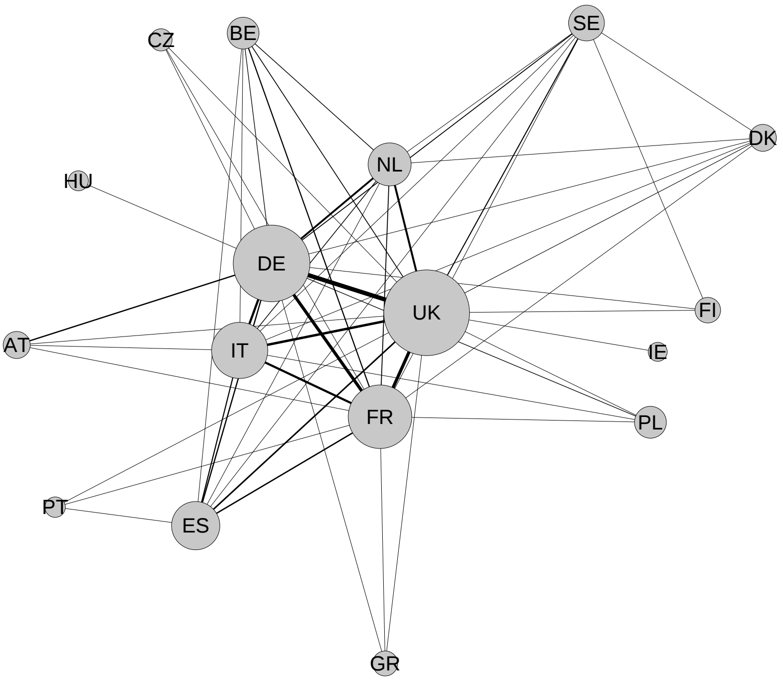


Figure 4. The excellency network between countries: links of type 1 (black) and 2 (light gray). The node size is proportional to the number of articles.

25% of both the total number of articles and the Top10. This is the backbone of the network.

3.0.2 NUTS3 Level

Although the above results show a global increase in both visibility and quantity of papers produced in both years, the distribution of article production is not homogeneous through all European regions. There are for instance “active” areas in each country, like the London area, or the Ile de France (around Paris). Although the largest fraction of NUTS3 are hotspots (60%), there is a large fraction of type 4 (29%) that have contributed in a large part to the increase in the number of publications.

More collaborative NUTS3 have a slightly higher visibility, see Figure 5: the percentage of Top10 collaborations has been plotted versus the number of inter-NUTS3 collaborations in a log-log scale. The behavior is the same for both years, the slope of the fitting line is consistently higher than zero, and increase between both years: thus, collaborating with a high collaborative NUTS3 has a higher probability of ending in writing a Top10 article.

In both years there are more international links than national ones. But the number of international links increases more (57% versus 24%). The total number of collaborations has increased from 5.2 10^5 in 2000 to 1.3 10^6 (almost doubled) in 2007. In both years, the total number of collaborations is equally distributed between national and international ones, thus indicating that although international links represent the majority, they are weaker than national ones. Moreover, the number of national Top10 collaborations has doubled between both years, but the international ones have almost tripled. On average, the probability that a collaboration produces a Top10 ar-

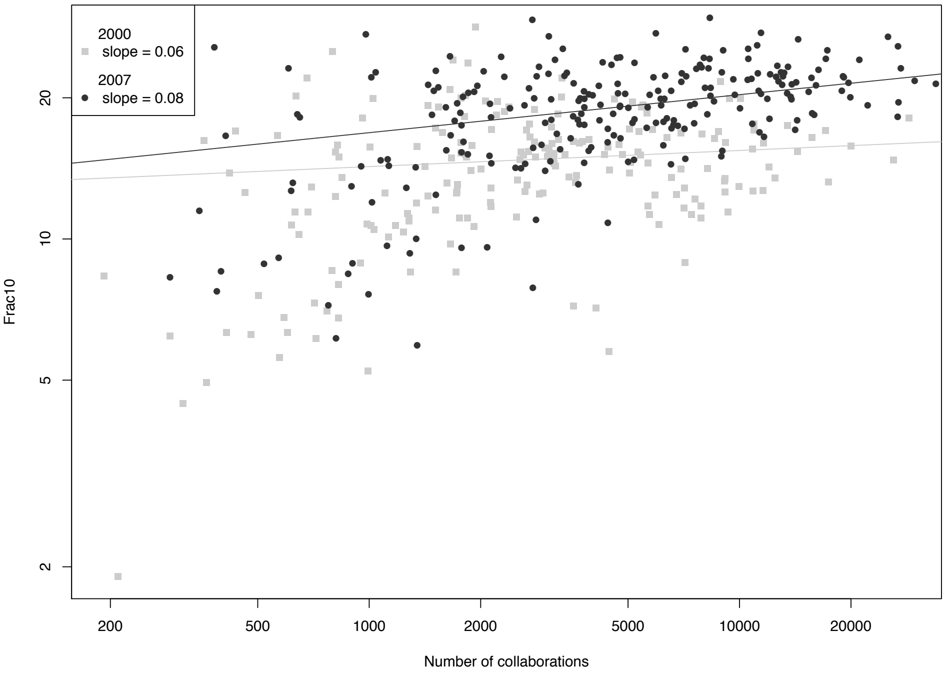
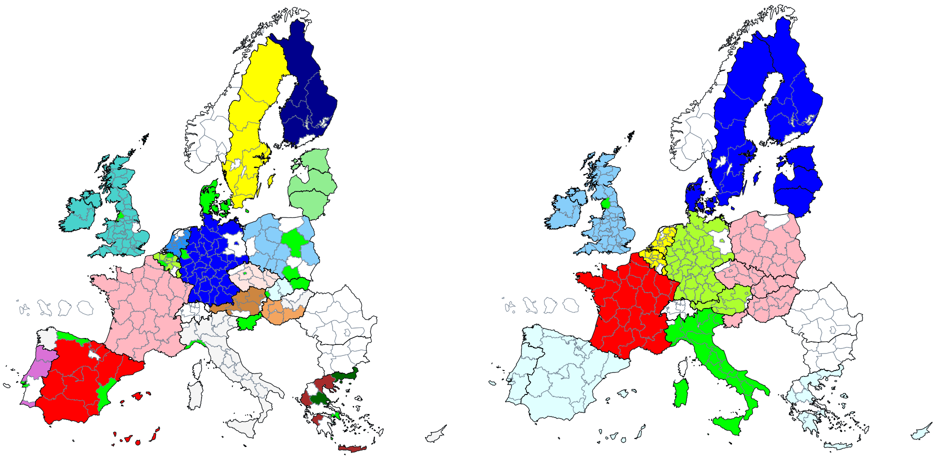


Figure 5. Log-log plot of Frac10 vs. number of collaborations articles. The line has been fitted on NUTS3 having at least 100 articles collaboration. The standard deviation is evaluated by linear regression anaysis.

ticle is 14% (12% if national, 16% if international) in 2000 and 20% in 2007 (14% if national 25% if international).

Distance plays an important role here. Using the Louvain algorithm [4], we have established a community structure of the NUTS3 network, see Figure 6. Communities of coauthorship have basically a national/linguistic character in 2000, mainly coinciding with entire countries (with the exceptions of the UK, Ireland and the Baltic republics). In 2007, however, most communities grow geographically to include neighboring countries, with the exception of Italy, France and Greece that still constitute single country communities. There are also large international communities: the area around the Baltic sea, the BENELUX, Spain and Portugal , and former Warsaw pact countries, that also have a long history of communication and trade.

The national characters of 2000's communities can be interpreted as an effect of the average weight for national link as compared to international ones. In 2007, the increase of the weight of national links is compensated by the large amount of international links. Contrary to expectations, NUTS3 of countries like France and Italy have a ratio between international collaborations and national ones less than 1 in both years. This means that collaborations are more national and only some NUTS are particularly active. On the other hand, countries like Austria, Portugal and Spain have considerably increased this ratio between both years, meaning that they are more prone to international collaborations. A gravity analysis as in [5] has shown, as expected, that being neighbors fosters collaboration.



(a) 2000

(b) 2007

Figure 6. Communities of coauthorship in 2000 (left) and 2007 (right). For readability, we have shown the communities at NUTS2 level.

The excellency network consists of all the NUTS3 and comprises 5.8 10^4 links. Almost two thirds of the links are hotlinks, they provide 80% of the total and also 80% of the Top10 collaborations. Almost two thirds of hotlinks are international, but, as stressed above, the majority of collaborations are national, while 60% of them are Top10 collaborations. On the other hand, for success links (20% of links), the number of international collaboration is almost three times the number of national ones, and the number of Top10 international collaborations are almost 4 times the number of national ones. These hotlinks and success links represent 84% of all links, 85% of all collaborations and also 85% of Top10 collaborations.

Figure 7 shows an excerpt of the excellency network: only nodes of with at least 5000 articles. Links tend to be preferentially established between hotspot NUTS (white) and reputationloss NUTS (grey). We define the "quality backbone" as the links where the production and the quality has particularly increased in both years, and consequently the probability of writing an article is much higher on this connections than others. Some countries have a well defined backbone:

- In France, the backbone is mainly formed by links with the Île de France (FR101, FR104, FR107), Rhône (FR714), Isère (FR716) and Marseille (FR824). Except Paris (FR101) and Essonne (FR104), all NUTS are of type 2 (focus on success).
- In the italian case, the backbone is the Milan Area (ITC45), Turin (ITC11), Rome (ITE43), Bologna (ITD55) and Padua (ITD36). The production in these NUTS has been focused on quality more than on quantity;
- In Spain, collaborations between Madrid (ES300) and Barcelone (ES511) dominate.
- In Germany, it is Berlin (DE300), Munich (DE212) and Hamburg (DE600).
- In the United Kingdom case the backbone is formed by collaborations with London (UKI11) from Bristol (UKK11), Oxford(UKJ14), Manchester (UKD31), Edinburgh (UKM25) and Glasgow (UKM34).

Note that English and French backbones are highly centralized. Moreover, these two centers are strongly collaborating and producing a large fraction of Top10 articles, thus acting as a bridge between both areas. We see as well that the nodes of the backbones are essentially hotspot or reputationloss, this is particularly striking if we

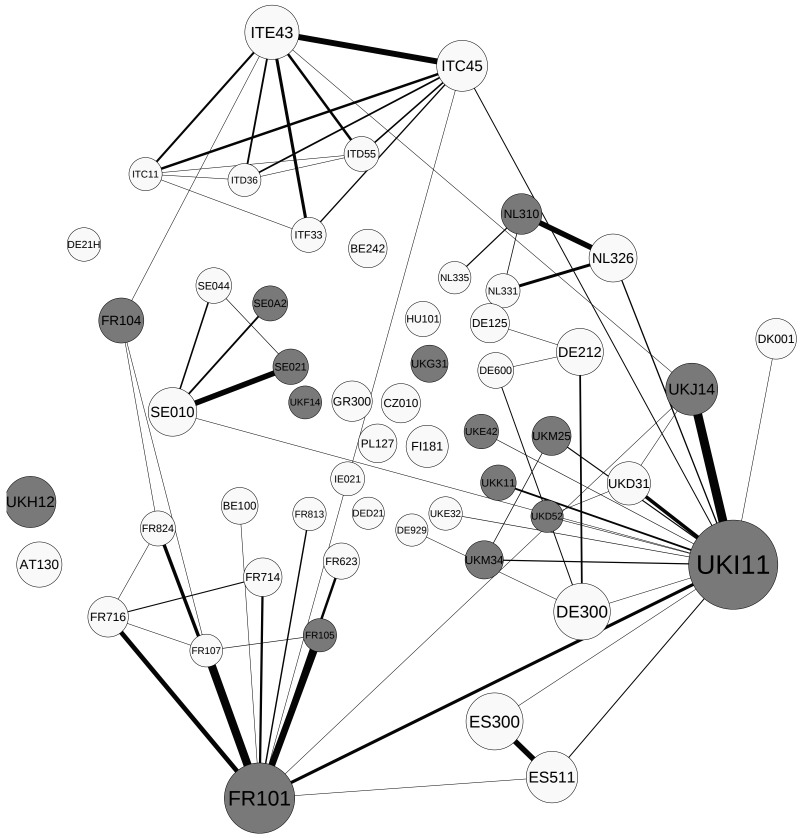


Figure 7. The excellency network. Black links are hotlinks (both the number of collaboration and the quality has increased between 2000 and 2007), grey ones are success links (the number of collaborations decreased but the quality increased). Link width represents the weight (number of articles). Node size is proportional to the number of articles. Node color reflects the type of nuts: hot spot white, reputationloss light gray.

consider the cases of London (UKI11) and Paris (FR101). As in the case of countries we evaluate the visibility variation v, for all the nodes in the backbones. Except for Essonne ($v = -58\%$) and Hamburg ($v = -11\%$), the visibility variation is always positive, in some case it has more than tripled (Isère, Rhône, Milan and Bologna).

In order to gain some insights on the effect of distance over collaboration visibility, we have compared the coauthorship network with a null model [11], where links are shuffled, preserving for each NUTS the number of collaborations and the Top10 fraction. In this way the null model incorporates no distance effect on the distribution of the links. Figure 8 shows the ratio between the number of collaborations for the real data and the null model (grey line) vs. distance, and the same ratio for the Top10 (black line). For collaborations of range less than 1000 km, the total collaboration ratio is larger than one, indicating that short range connections are stronger (and not occasional). For longer ranges, the ratio is almost one, indicating an effect of the distance on the number of collaborations. For distances above 2000 km, random fluctuations appear. In the case of Top10 collaborations, the ratio is always greater than one. The 1000 km limit is interesting in itself, since the largest number of connections above this range are international, thus strengthening the idea that international collaborations have a higher visibility.

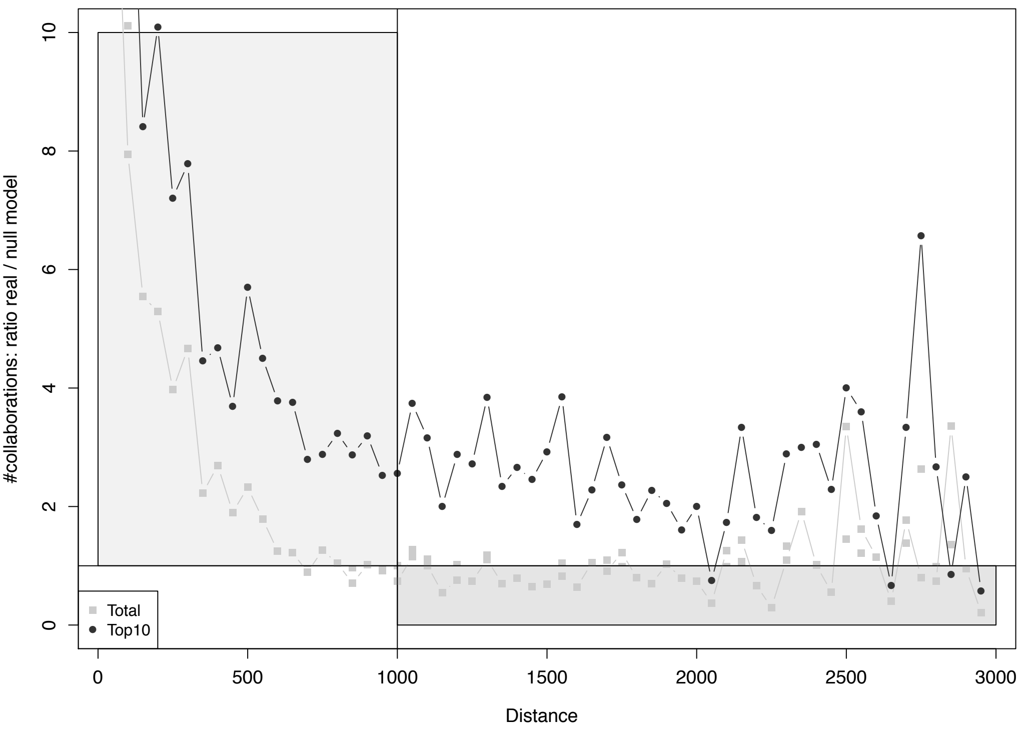


Figure 8. Comparison between empirical data and null model network. The gray line represent the ratio between the empirical collaboration at a certain distance and the null model one. The black line represents the ratio for the Top10 collaborations.

We finally draw on on Figure 9 the average Frac10 weighted by the number of collaborations at that range. On average, the fraction of article in Top10 is higher in 2007 than in 2000 at almost all ranges. After an initial increase until 1000 km, the distribution becomes more stable and eventually oscillates for long ranges.

Although this analysis is not a conclusive one, it shows that geographical distance can have an effect on article visibility. Moreover this is an another confirmation of the positive effect of international collaborations on article visibility.

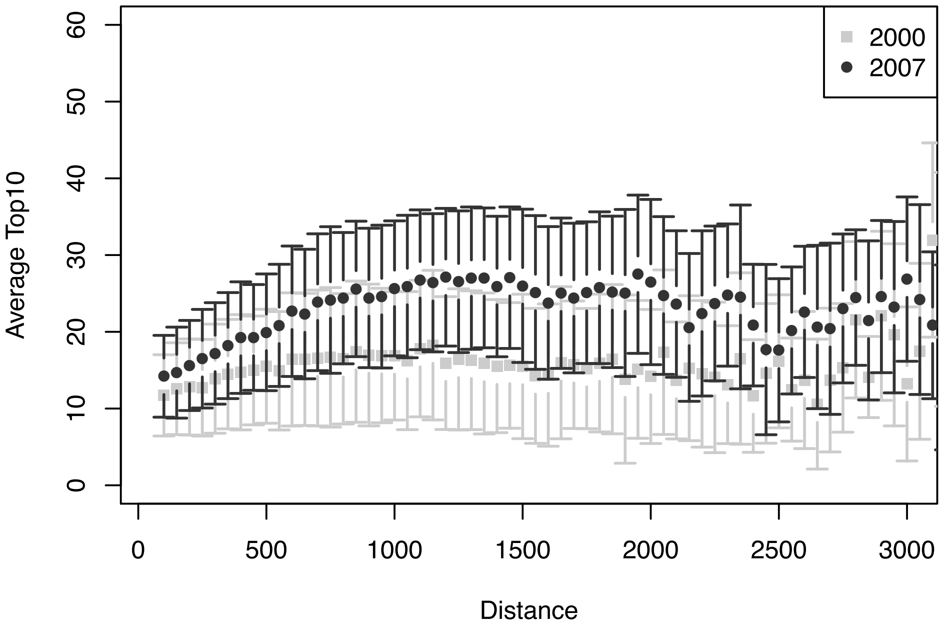


Figure 9. Frac10 as a function of distance.

4 Conclusion

In this article we have studied the impact of geography on article visibility, and presented a coauthorship network. For this, we used a dataset from the OST, containing all articles written in 2000 and 2007 with at least one author affiliated in Europe. The study has focused on the evolution of the qualitative characteristics of the network and on the geographical distribution of the coauthorships. Since different geographical units can be defined as nodes of this network (cities, countries, regions, NUTS3), this is an example of multi-scale analysis, where the type of aggregation of sub-units can hide information about heterogeneities. The longitudinal analysis of the network, comparing both years of data, has shown an increase in both quantity and visibility of the articles produced in Europe. At the country level, all countries have increased (at different rates) both the Top10 and the total productions. However the analysis at the microscopic level (NUTS3) has shown more heterogeneities. Between 2000 and 2007, there has been an important increase in the number of connections, in particular the international ones. Coauthorship analysis has shown that, while in 2000 virtually all communities where national, in 2007 some of them have grown to include neighboring countries.

The longitudinal study has noticed that strongest *hot links* have national character, and thus constitute the *quality backbone*. The different shapes of the national backbones vary from country to country: they are highly centralized in France and the UK but more dispersed in Italy, as could be expected. The strength of the connection between Paris and London acts as a bridge between France and the UK and also between Southern and Northern Europe, while links through Germany bridge Western and Eastern Europe. International links have a much higher visibility.

In order to better understand the effect of distance on article visibility, a detailed analysis of the different disciplines would be necessary, since the use of big experimental apparatus could contribute to creation of large international collaborations (an example of which are CERN publications).

Acknowledgements

A.A. and P.J. work is supported by DynaNets. DynaNets acknowledges the financial support of the Future and Emerging Technologies (FET) program within the Seventh Framework Program for Research of the European Commission, under FET-Open grant number: 233847

References

1. M.E.J. Newman, Proceedings of the National Academy of Sciences, **98**, (2001) 404-409.
2. M.E.J.Newman, Proceedings of the National Academy of Sciences, **101**, (2004) 5200-5205.
3. C. W. Matthiessen, A. W. Schwarz and S. Find, Urban Studies, **47**, (2010) 1879-1897.
4. V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment, **10**, (2008) 8-20.
5. P. Jensen, A. Apolloni and J.-B. Rouquier, *Donnees urbaines volume 6* (Economica, 2011) 279-288.
6. L. Bornmann and A. Plume, , Journal of Informetrics, **5**, (2011) 695 - 697.
7. J. Hoekman, K. Frenken and F. van Oort, KITeS Working Papers, **214**, (2008) 1-22.
8. K. Frenken , S. Hardeman and J. Hoekman, Journal of Informetrics, **3**, (2009) 222 - 232.
9. K. Frenken, Economic Systems Research, **14**, (2002) 345-361.
10. J. Katz, Scientometrics, **31**, (1994) 31-43.
11. J.-P. Onnela, S. Arbesman, M.C. Gonzalez, A. L. Barabsi and N. A. Christakis, PLoS ONE,**6**, (2011) e16939.
12. K. Frenken,Frank van Oor and Roderick Ponds, Papers in Regional Science,**89**, (2010) 271-351.
13. T. Opthof and L. Leydesdorff, CoRR, (2011).
14. M. Bonitz, E. Bruckner and A. Scharnhorst, Scientometrics,**44**, (1999) 361-378.