RIERC Regional Innovation and Entrepeneurship Research Center

WORKING PAPER SERIES

2016-03

How to get from the periphery into the core? The case of co-authorship network evolution in neuroscience

Tamás Sebestyén, Orsolya Hau-Horváth, Attila Varga

Regional Innovation and Entrepeneurship
Research Center

Regional Innovation and Entrepreneurship Research Center
Faculty of Business and Economics
University of Pécs

H-7622, Pécs Rákóczi str. 80.

Phone: +36-72-501-599/23121

www.webcím.hu

How to get from the periphery into the core? The case of co-authorship network evolution in neuroscience

Tamás Sebestyén

MTA-PTE Innovation and Economic Growth Research Group

University of Pécs, Faculty of Business and Economics

Orsolya Hau-Horváth

MTA-PTE Innovation and Economic Growth Research Group

University of Pécs, Faculty of Business and Economics

Attila Varga

University of Pécs, Faculty of Business and Economics

MTA-PTE Innovation and Economic Growth Research Group

Abstract

It is well documented that geography and networks coevolve. We use a wide database of coauthorship which allows for the investigation of this phenomenon. The geographical position of an author has an effect on performance and network position. Looking at the career paths those authors perform better in these two respects which change geographical location. We detect a marked role for double affiliation (being at the center and the periphery at the same time) both for performance and network position. This finding is reinforced by a regression analysis.

Keywords

Co-authorship networks, neuroscience, ENQ, network position, geographical location

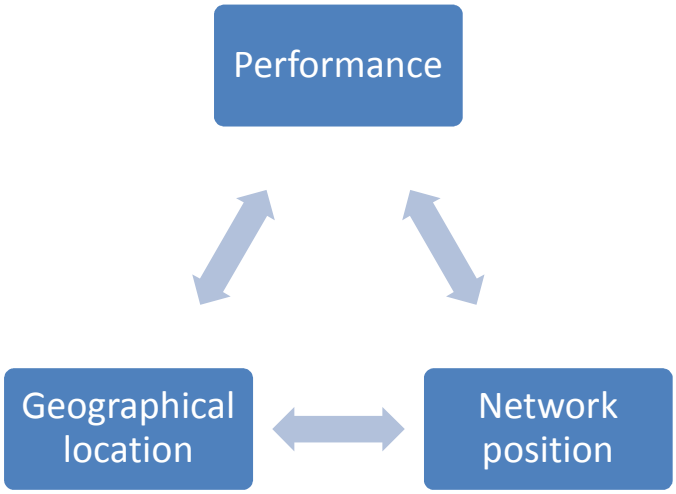
1. Introduction

The literature on co-authorship networks reveals several insights into the structure and working of scientific collaboration, but there are some focal issues these studies concentrate on. One issue is the relationship between collaboration and performance which latter is typically measured by citation counts. There is evidence on the positive effect of network centrality (Acedo et al., 2006; Abbasi et al., 2011; Yan and Ding, 2009; Abbasi et al., 2012), closeness centrality (Yan and Ding, 2009), betweenness centrality (Yan and Ding, 2009; Abbasi et al., 2012), PageRank centrality (Yan and Ding, 2009), as well as tie strengths (Abbasi et al., 2011) and efficiency á'la Burt (Abbasi et al., 2012).

Interestingly Abbasi et al. (2011) finde a negative effect for eigenvector centrality while Ahn et al. (2014) show that links to high-reputation universities also contribute positively to performance.

Another question, following directly from the previous one is if collaboration patterns determine performance, what determines collaboration patterns? A wide range of studies emphasize the role of network distance (Fafchamps, 2010), technological proximity (Cunningham and Werker, 2012), geographical proximity (Cunningham and Werker, 2012; Hardeman et al., 2012) institutional proximity (Hardeman et al., 2012), academic excellence and informal communication (Jeong et al, 2011) and also similar affiliational background (Rodriguez and Pepe, 2008).

These studies mainly focus on the topological features of the network although some pay attention to geography. Our goal in this paper is to more explicitly take into account the role of geography on the global scale when examining the relationship between network properties and scientific performance. Our main intuition is that network position, geographical location and scientific performance are co-evolving over the career of researchers and one can not analyze one or two of these concepts without taking into account the other(s).



1. Figure – The conceptual framework

In this paper we introduce a database primarily built for the analysis if this trial interrelationship. Although there is no room to tackle this issue in its entirety we present some first results with respect to the role of geographical position. The paper is structured as follows: in Section 2 we describe the database and introduce the measurement tools we employ. Section 3 contains some descriptive statistics on the relationship between performance, network position and geographical location. In Section 4 we show the results of a regression analysis which focuses on the role of geographical location and performance in explaining network position of scientists. Section 5 summarizes the findings of the paper.

2. The dataset and measurement

This research builds on the database of Scopus, containing information on a wide range of scientific periodicals. We examined the possible use of different publication databases but finally decided to use the Scopus database as this is the largest available dataset on publications and citations indexing as many as 21 thousand different products of 5000 international publishers, including 20 thousand

refereed journals.[1] In the following sections we describe how the final dataset under examination was built from this huge amount of data and then introduce the measurement of the concepts mentioned in the introduction using the data available in the database.

2.1. Data preparation

As the present research focuses on the field of neuroscience research, we restricted our search in the database for those journals which belong to the field neuroscience. Scopus categorizes the different journals with respect to the scientific field they belong to and one source may be categorized under several different fields. We retrieved all journals from the database which were categorized under neuroscience, meaning 498 journals in total.

The main focus of this research is on the best performing authors in neuroscience, so we should further restrict the database with respect to authors in order to be left with really the top performing scientists in the field. This restriction was done in two steps: first, we restricted the journals under consideration, focusing on 'top' journals. Second, we restricted the authors for 'top' authors with the best publication performance.

In order to narrow down the list of journals, we took the SJR (SCImago Journal Rank) indicator as a basis.[2] The data we retrieved from Scopus contained information on 2008, 2009 and 2010 SJR scores, and we used the average of these three years for journal selection. We decided to exclude those journals which do not have a positive SJR score in these three years: this means 376 journals out of the 498 in the field of neuroscience. Then we ranked the journals according to their average SJR scores starting from the top journals and calculated the relative cumulative SJR score for each rank: e.g. a relative cumulative score of 15% at rank 5 means that the first 5 journals account for 15% of the sum of SJR scores for all journals. We decided to restrict our sample of those journals which account for 50% of the SJR scores which meant that we retrieved the first 57 journals in the field of neuroscience.

After the restriction of the sample to these journals, we retrieved the following information for all documents (publications) in these journals from 1974 to 2014,[3] available from the Scopus website:

- Last name of the authors and the first letter(s) of their first name(s)
- The title of the document (article)
- The title of the source (journal) in which the document is published
- The country of the affiliation of the authors (more than one affiliation is possible)
- The affiliation of the authors (more than one affiliation is possible)
- The year when the document was published
- The number of citations received by the documents up to the date of the data retrieval (December, 2013)

[1] See: http://www.elsevier.com/__data/assets/pdf_file/0007/148714/scopus_facts_and_figures.pdf).
[2] The SJR indicator is used for journal ranking which calculates the weight of the different journals according to the weight of those journals which cite that journal. See González-Pereira et al. (2010) for details.
[3] Although the time of data retrieval was December 2013, the database contains information on some 2014 publications as well which were supposedly dated forward in the respective journals. However, we excluded 2013 and 2014 from the analysis as the information for these years must not be complete.

This restriction of the journals to the top 57 periodicals still contain around 340 thousand documents (publications) and altogether 1.5 million publication-author pairs. The time span of the database is quite large: we have publications from 1974 to the date of the data retrieval which is January, 2013. This time window spans the entire life cycle of the scientific field under question as the first seminal publications in neuroscience date back around the time of our first year.

A large task after the data retrieval was to clearly identify authors to follow their publications. We have to be aware of misspelled names as well as the same author publishing under different names and also authors with the same name. We employed different methods to come up with an acceptable identification of authors. Due to the huge amount of data we employed algorithmic methods which are relatively fast but may leave some inconsistencies in the data. First, a character-distance method was used to identify possibly identical authors with similar names. Then, we used the affiliation data for the authors to search for further similarities in different author names and differences under the same author names. Finally we attached a unique identifier for each author – the dataset contains 370 thousand author identifier altogether.

If we are to focus on top authors, it is clear that having 370 thousand in the sample is simply too much, so we employed a second round of restriction narrowing down the sample of authors to the 'top' authors who publish the most according to the available data. This restriction is intuitively driven by the observation that out of the 370 thousand authors more than 50% has only one publication under the whole time span of the dataset, and more than 75% has at most 3 publications. Finally we restricted the sample of authors on those scientists who have more than 40 publications in total. This threshold was obtained by focusing on the top 1% of the authors. The logic of the restriction was similar to that used for the journal selection. The authors were ranked according to the number of their publications, starting from the highest publication record. Then we calculated the cumulative publication share for each author/rank. A cumulative publication share of 0.1% at rank 372 e.g. means that the first 372 authors account for 0.1% of the total publications in the dataset. Using this method we put the threshold where the authors in the list account for 1% of the total publications – this corresponds to 41 or more publications and the first most publishing 3838 authors, which is cc. 1% of the authors after identification in the 57 journals.

Finally, we have a dataset for analysis which contains the top scientists in neuroscience (with respect to their publication records), their publications in top journals in the field (with respect to SJR scores). This means 3838 unique authors and 57 journals.

2.2. Measurement and final database

Using the available data outlined in the previous section, we try to analyze the interrelationship between geographical location, network position and scientific performance. In order to quantify these concepts, we apply the following measurement techniques using the data available from the database we constructed.

Geographical location of an author

The database contains two sources of information on the geographical location of the authors. First, we have a country code for each author and for each publication – one author may have several countries at the same time if he or she has multiple affiliations. Second, we have data on the exact affiliation which would allow us in principle to provide a detailed (sub-country) identification of the authors' location. However, the affiliation data is available in character format, and it would need

vast clearing to bring this information into a usable format. On the other hand, even using countries as the basis for geographical localization would overwhelm the analysis, we decided to group the countries into two groups: the geographical center and the geographical periphery. The categorization is based on the occurrences of the countries as author location at the beginning of the sample: those countries are classified as belonging to the geographical center where author published the most. Not surprisingly, these countries correspond to the economically advanced countries: USA, EU15, Japan, Canada, Australia, Switzerland, Norway and Israel. Of course, this is only one possible classification and some countries made significant development in their publication performance in the field – e.g. South Korea and China seem to be catching up with the core countries. On the other hand, the group of countries where the most publications are recorded remains quite stable over the sample, and for this reason we retain the classification based on the initial periods for the whole sample.[4]

Using this classification we establish two measurements for the geographical location as described below:

- *Discrete location*. According to this measurement an author is characterized into three possible categories: (i) having affiliation only in countries classified as the geographical center; (ii) having affiliation only in countries classified as the geographical periphery; (iii) having affiliation in both regions.
- *Continuous location*. According to this measurement an author is characterized by a value between 0 and 1 which measures the percentage of his or her affiliations which belongs to the geographical center. The higher this number, the more exclusively an author is working in the geographical center.

Network position of an author

Using the information retrieved from the Scopus database we are able to build a co-authorship network of neuroscientists for all years of the sample. The network we use is a weighted network where the weight of a co-authorship tie is based on the number of publications the two authors had together. As we have a long time span, the co-authorship network is constructed for each year separately. In formal terms let $a_{i,j,t}$ denote the weight of co-authorship link between authors i and j in year t. This weigh is simply the number of publications which authors i and j co-published in year t. Having this raw view on annual publication records can be misleading with respect to the duration of interpersonal links: links are only recorded for the year when the publication is published. It is intuitive to think, though, that several years of work precedes this publication, so the interpersonal relationships exist even before the publication date. In order to account for this bias, we use a 5-year moving aggregation. This means that in the final network data, used for the analysis later on, the link weight between two authors is the sum of the raw link weights in the given year and the four preceding year. Formally, we define $b_{i,j,t} = \sum_{\tau=0}^{4} a_{i,j,t+\tau}$ and $b_{i,j,t}$ is used in the analysis – so the link weight of a given year is the sum of co-publications in that year and the four consecutive years.

Using this network data we can calculate measures which capture the authors' position in the co-authorship network.

[4] There may be attempts to include a varying classification where countries can join and leave the geographical center but there are some methodological problems to be addressed once we are to operationalize this: what threshold to establish, what to do with countries which temporarily switch between the classifications, etc.

We operationalize network position using two different measures as described below.

- *Coreness profile*. Following the method described in Della Rossa et al. (2013) we assign a value ranging from 0 to 1 to each author which reflects the extent to which the given author belongs to the core of the network. The higher this value, the more connected the author is and the lower, the more peripheral he or she is.
- *ENQ index*. Following the method described in Sebestyén and Varga (2013a, 2013b) we assign a value to each author which captures the quality of available knowledge from the network of co-authorships.

The difference between the coreness profile and the ENQ index is that while the former takes into account the connection structure of the network, the ENQ index also accounts for the knowledge level of the direct and indirect co-authors. On the other hand, while the coreness profile is normalized to the interval between 0 and 1, the ENQ index is unbounded ad driven by the total knowledge in the network.[5]

Performance of an author

The third concept we intend to measure in this study is the scientific performance of the author. We give three possible measures for the performance of an author, based on the information available from the database:

- *Publications*. We simply count the number of publications of an author in a given year. This is the most common performance measure.
- *Citations*. We simply record the number of citations obtained by a given author. Although this is also a widely used measure of scientific performance, the limitations of our database call for some attention when using these records. The available information is on the citation count of a paper up to the point of data retrieval. We can then provide an annual measure for all authors: the citation value of an author i in year t is the number of citations on the papers published by author i in year t (up to the date of data retrieval). However, these counts must be biased at least for two reasons: (i) earlier publications have more time to accumulate citations, so citation counts of younger authors with most of their papers in recent years are biased downwards; (ii) citation counts even for one single author are not necessarily comparable for the same reason: older publications may accumulate more citations.
- *Citations over publications*. We simply divide the number of citations an author received with the number of paper he or she published in a given year. This can be regarded as an efficiency measure: what is the impact an author can reach through one publication. However, the problems of the citation counts are also valid in the case of this measurement.

The final database

Once these measures are calculated, we assembled the final database for analysis. As we have many observational units (authors) for many years, we can render the data into a panel database. However, in order to make comparisons between authors, it seems to be more interesting to render

[5] As it is described in Sebestyén and Varga (2013a, 2013b), the ENQ index requires the opreationalization of the knowledge levels of the authors. In this paper we use the cumulated number of publication as a proxy for indivdual knowledge levels.

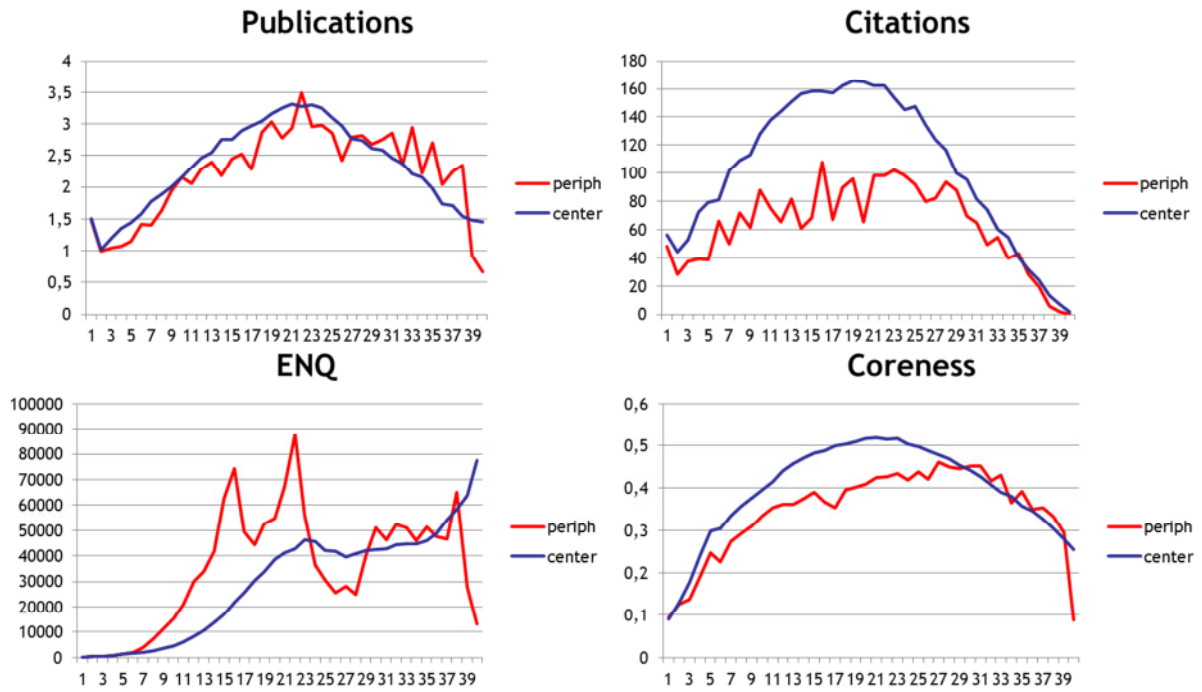
the panel database on a career-year basis rather than on a calendar-year basis. This means that we take each author, record the first year of his or her activity in the database (proxied by the first publication) and then this year is going to be his or her first year in the database. This way the database contains in the first period the first career-year of all authors and so on. Although the personal comparison is obtained here, the panel becomes unbalanced as there are a few authors whose life cycle span almost the whole period (around 40 years) but many of them have longer or shorter careers. Another source of bias is that although some careers are full in the sense that we have early and late career records for these individuals, but many careers are truncated as these authors are still in their early or middle career years at the date of data retrieval.

3. Descriptive statistics

In this section we provide and discuss some descriptive statistics of the data established in the previous sections. The general logic behind these statistics is that we split the sample of authors according to some characteristics (e.g. locating in the geographical center or in the geographical periphery) and then see if the different subsamples have different patterns with respect to some variables (e.g. performance, network position) over the career years. First we show some analysis with respect to the role of geographical location (i.e. when subsamples are made according to the geographical location) and then on the role of network position (i.e. when subsamples are made according to network position).

3.1. The role of geographical position

In our very first attempt the sample of authors are divided into two groups according to their life-time geographical position. We took the *continuous location* measures – each author has one value for all career years – and calculated the individual averages over each author's lifetime. These values reflect that to what extent an author belong to the geographical center or periphery over his or her lifetime. We then split the sample between 'mainly center' and 'mainly periphery' authors, the former group containing those authors whose average location value is above (or equal to) 0.5 and the latter group containing those whose value is below 0.5.



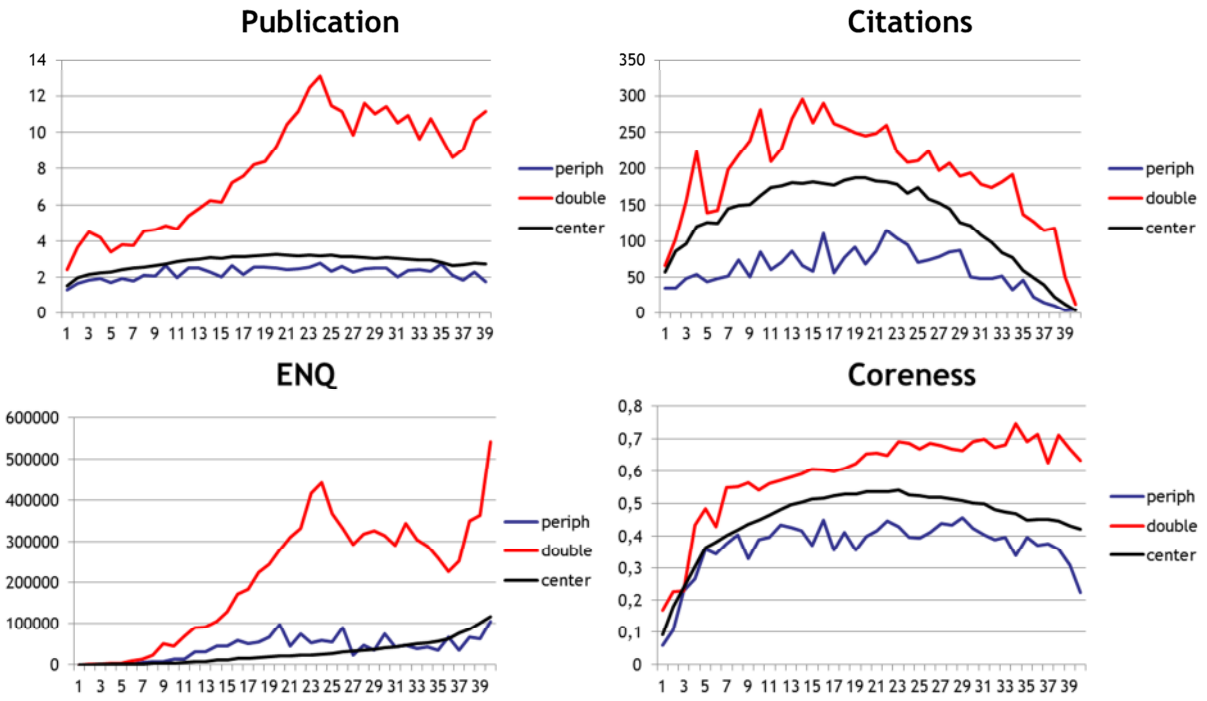
2. Figure – The role of main lifetime geographical location on performance and network position

Figure 1 collects the data with respect to this analysis. The horizontal axis on each panel corresponds to career-years. The blue lines represent the average publication, citation counts and ENQ and coreness values for those authors who on average belong to the geographical periphery over their lifetime and the red lines represent the average values of those authors who on average belong to the geographical center over their lifetime. In each panel the title defines the value which is measured on the vertical axis.

Some observations are straightforward here. First, the line representing authors locating mainly in the geographical center is smoother. This comes from the fact that the size of this subsample is much larger than the periphery subsample: the authors in our database typically locate in the geographical center (numbers here!). Second, except for the ENQ index, all measures exhibit a reversed-U shape which nicely shows that the researchers in the sample are the most active (have more publications and better network positions) in their mid-career while during their early and late career they are less active in publications, publications in these life stages receive less citations and they move out from the core of the network. On the other hand, the ENQ index shows a different path: increasing throughout the time window which is due to the fact that this index is driven by the knowledge level measured by cumulated publications which increase by definition. On the other hand, observing the contrast between the coreness profile and the ENQ index we can argue that although the authors move out from the network core in their late careers they manage to maintain important connections which link them to the most knowledgeable partners.

With respect to the difference between the two subsamples (authors mainly in the geographical center and geographical periphery) we can also see some important results. First, it seems that the authors' overall or typical geographical location does not make a difference in publication counts – on the other hand, those authors who belong mainly to the geographical center have much better citation performance in their mid-career. Second, geographical location also differentiates with

respect to network position. If we take the coreness profiles, we see that authors in the geographical center are better positioned in the network in their mid-career, but in the early and late careers the difference is not significant. Third, if we look at the ENQ index, we see the reverse: in the mid-career phase those authors have typically higher ENQ index who mainly belong to the geographical periphery. This difference may come from the fact that although these authors are not in the core of the network as shown by the coreness profile, they manage to establish and maintain important links towards knowledgeable partners which compensate for the less favorable network position. This result may show why these authors, in spite of their less favorable geographical position are able to become one of the top authors in the field.



3. Figure – The role of current geographical location on performance and network position

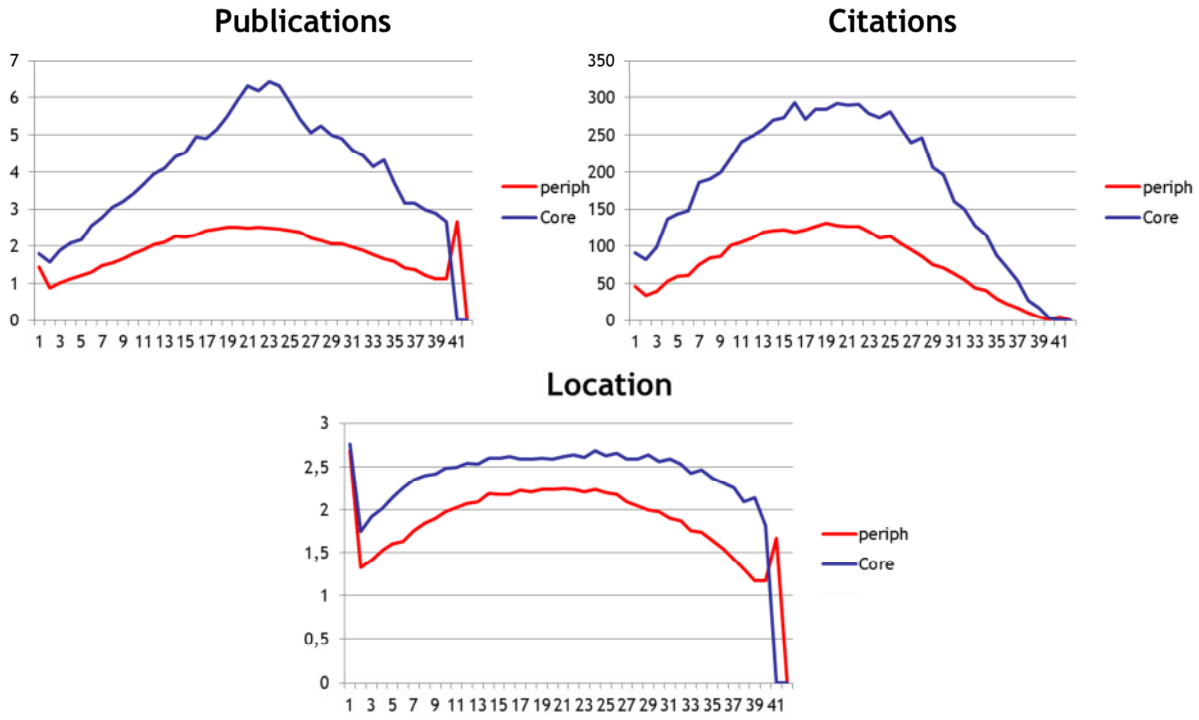
In Figure 2 we show a similar analysis, but the differentiation is now done on the basis of the current location of the authors. While in the previous analysis each author had one category according to which he or she belongs to the geographical center or periphery on average over the lifetime, now each author has a category for each year of his or her career depending on the affiliations he or she has in that given year. In contrast to the previous analysis where we used the continuous location measure, now we base the analysis on the discrete location measure. According to this all authors (in all years) are classified into three categories: affiliated only in the geographical center (black lines), affiliated only in the geographical periphery (blue lines) or affiliated both in the center and the periphery (red lines). The main features of the data are also reflected here as in Figure 1: curves are inverse-U shaped except the ENQ index, the majority of the authors are typically affiliated in the geographical center in all career-years and authors affiliated in the geographical center outperform periphery authors in citations, coreness and also in publication (this was not the case with the main geographical location). What is striking that authors with double affiliation (both in the geographical center and geographical periphery) even outperform those only affiliated in the geographical center, and this is true for all four indicators and for almost all career years. In the case of publication, those with double affiliations do not seem to decrease publication activity even in the late years of their

career as 'single-affiliated' authors seem to do. These results show that moving between the geographical center and periphery is able to positively influence network position and publication performance especially if the former affiliation is maintained.

It is also interesting to see whether the initial geographical location of the authors makes any difference in their performance and network position over their career. We also checked this question by grouping the sample according to the initial geographical position of the authors. Without presenting the diagrams we note that the initial geographical location does not seem to make any difference in performance and network position – the role of the double affiliation is also not present in this case.

3.2. The role of network position

In the previous section we looked at the role of the geographical location of authors on their performance and network position. In order to further examine the interrelationship between the three concepts as introduced in the introduction, now we look at how being at different network positions affect performance and geographical location. In other words, the sample of authors is now split between those authors who are in the network core and those who are at the network periphery and we examine whether there is any significant difference between the two groups with respect to their geographical location and performance.



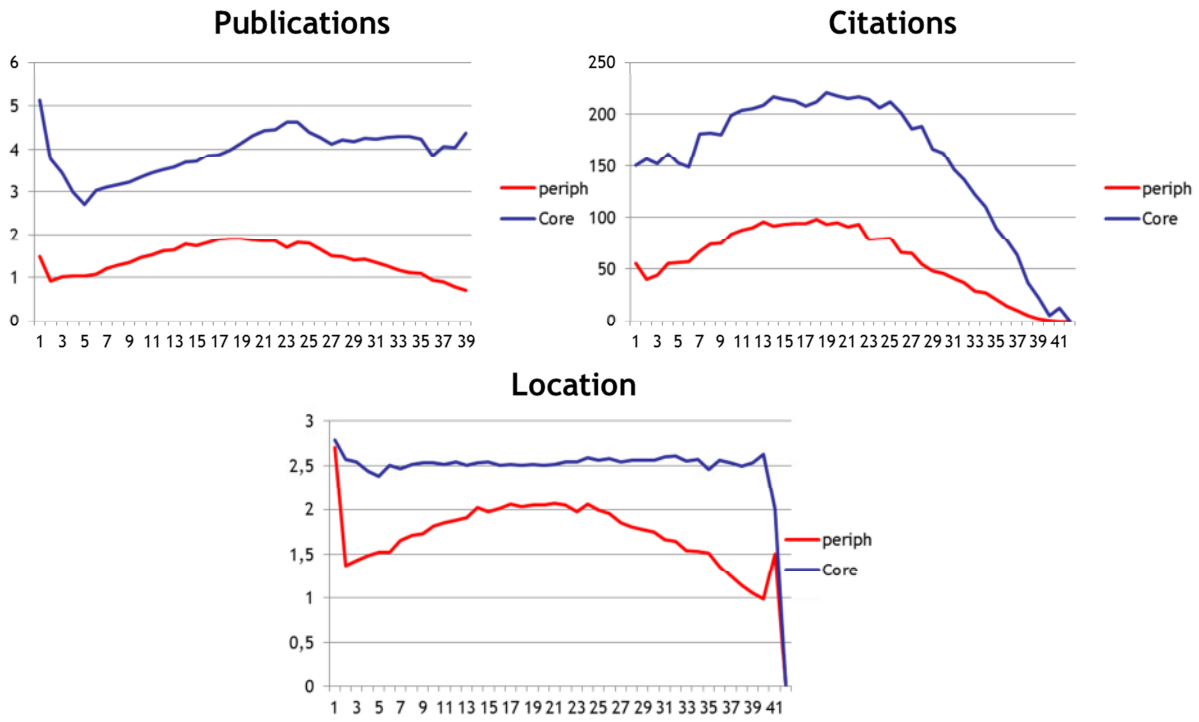
4. Figure – The role of main network position on performance and geographical location

In Figure 3 the sample is split along life-time network position. We took the coreness profile of each author for every year and calculated the average coreness value for all authors over their lifetime. Then all authors with a 0.5 or higher value were classified as belonging to the network core and

those with lower than 0.5 were classified as belonging to the network periphery as their main or average network position.[6]

The upper two panels of the figure read analogously to the previous figures. The red line shows the average publication and citation values of authors belonging to the network periphery on average while the blue lines that of the authors in the network core. These figures reflect again the inverse-U shape as before and show a significant difference in performance in favor of the authors who maintain their core network position over their lifetime. It is interesting to see that the relative difference between the two groups of authors is around two-fold overall the authors' career and in both performance measures.

The bottom panel was constructed using the discrete location classification of authors. We assigned a value of 1 to those authors who are affiliated only in the geographical periphery in a given year, a value of 3 to those authors who are affiliated only in the geographical center and a value of 2 to those authors who have affiliation in both the geographical center and the periphery. As the two lines represent average values for the two sub-samples (network core and network periphery authors), the values on the vertical axis reflect the extent to which authors in the given subsample are located in the geographical center, periphery or in between on average. In this panel we observe, similarly to the other two panels, that authors who spend the majority of their career positioned in the core of the network are those who are located mainly in the geographical center.



5. Figure – The role of current network position on performance and geographical location

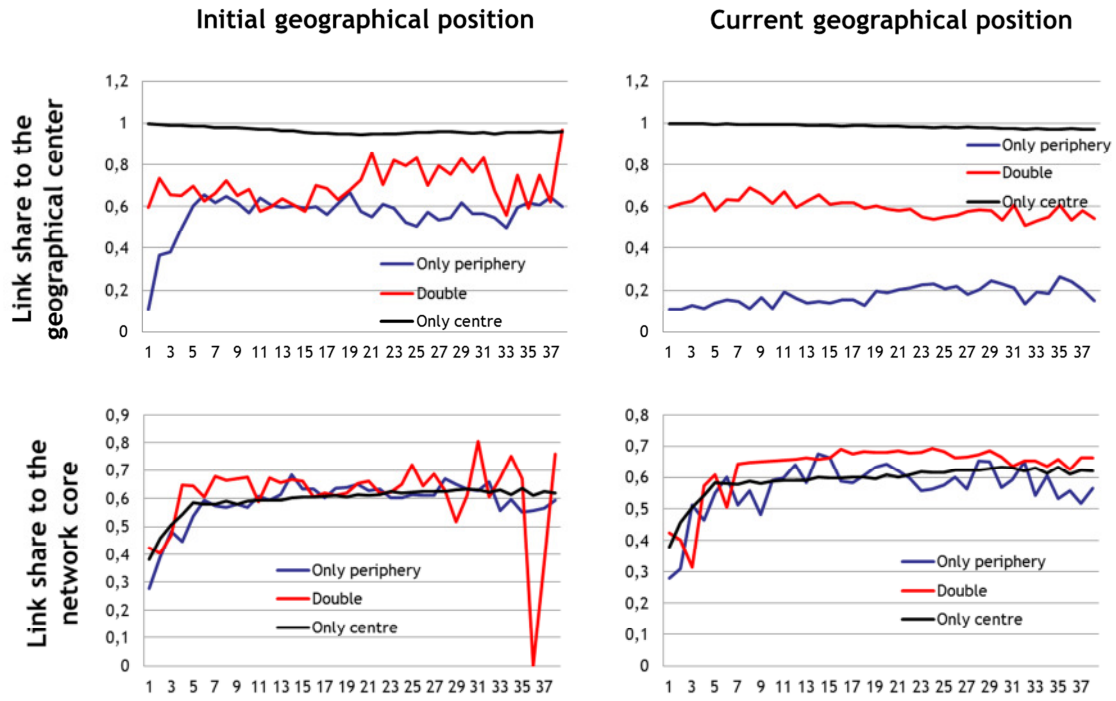
Figure 4 reflects the differences in network position but instead of the average lifetime position, now the current, annual position is the basis for splitting the sample of authors. The results here

[6] Although the co-authorship network is scale-free with a power law degree distribution, the corness values are scattered symmetrically around 0.5.

reinforces those obtained from Figure 3: a core position in the co-authorship networks positively affects performance and better connected authors are more likely to dominantly be located in the geographical center. According to Figure 4 this is not only true on average over the lifetime of the authors but also on an annual basis.

3.3. Link formation processes

In the previous sections we focused on the position of the authors in the network and their geographical location as well as their performance. In what follows, our focus is not only on the position of the authors but where they links point and how this is related to their position in the network and their geographical location. In order to operationalize this, we calculated for each author and for each year the share of links which points to the geographical center and also the share of links which point to the network core.



6. Figure – The role of initial and current geographical location int he share of links pointing to the geographical center and the network core

Figure 5 provides some insight into the relationship between geographical position and the direction of links with respect to network and geographical positions. The figure reads similarly to the previous ones: each panel shows average link share values for three subsamples, differentiated along the discrete geographical location measure. The upper two panels have the link share pointing to the geographical center on the vertical axes while the bottom two panels have the link share pointing to the network core on the vertical axis. On the panels in the left column the sample of authors is split according to their initial geographical location whereas on the right panel the current annual location is the basis of the subsamples.

With respect to the link share to the network core the first observation is that there is no significant difference between the subsamples irrespective of the basis of the categorization. All authors start at the relative periphery of the network, around 30-40% of their links pointing into the network core and then relatively rapidly they develop their position with around 60% of their links pointing into

the network core already at the end of the first decade of their career. We do not see any role for the initial geographical location and for the current geographical location in this sense.

On the other hand, the picture is very different when we focus on the geographical arrangement of connections (upper panels). The first striking point is that in spite of the fact that we did not find any effect of the initial geographical location on performance and network position (see section 3.1.) and also on the share of links pointing into the network core (see the bottom-left panel of Figure 5), there is a marked effect of initial location with respect to the share of links pointing to the geographical center. It is interesting to see that those authors who start from the geographical center have links which almost exclusively point to the center. There is a slight change in this over the career years but throughout their life time these authors are connected almost exclusively to the geographical center. On the other hand, authors starting at the geographical periphery have a very limited amount of links toward the geographical core but they increase this share quite rapidly. However, these authors retain their connectedness to the geographical periphery as their link share increases to around 60%. Authors starting with double affiliation move somewhere in between but due to the low sample size in this case we can not argue that there is a significant difference between authors starting from the periphery and with double affiliations.

The upper-right panel shows how the current position of the authors in a given year affect their share of links towards the geographical core in the same year. There is a very clear difference between authors affiliated only in the geographical center who almost exclusively maintain links with other authors also affiliated in the geographical center, while the case is reversed for authors affiliated only in the periphery. It is nice to see that double affiliated authors have a quite balanced network position with around 60% of their links pointing into the geographical center. These differences between the three subsamples do not change with the age of the authors.

4. Regression analysis

In the previous sections we heavily focused on some descriptive facts provided by our dataset about the interrelationship between network position, geographical location and scientific performance. In this section we use the panel structure of our dataset in order to carry out some regression analysis where we try to explain the evolution of network position with the other two concepts, namely geographical location and performance.

We employ a dynamic panel specification of the following general form:

$$NWPOS_{i,t} = \alpha + \delta \cdot NWPOS_{i,t-1} + \beta_1 \cdot GEOLOC_{i,t} + \beta_2 \cdot PERF_{i,t} + u_i + \varepsilon_{i,t} \qquad \text{(EQ1)}$$

where $NWPOS_{i,t}$ is a proxy for the network position of author i in career year t, $GEOLOC_{i,t}$ is the geographical location of author i in career year t while $PERF_{i,t}$ is a proxy for the performance of author i in career year t. u_i is an observation (author-) specific fixed effect while $\varepsilon_{i,t}$ is an observation-specific error term. Using this formulation we can examine whether geographical location and scientific performance affect the improvements in network position: the lagged network position on the right hand side controls for the persistence in network position over time, and also drives the regression to measure the effect of the other explanatory variables on the change in network position compared to the previous period. EQ1 is a general form which is operationalized with different measurements. Network position can be measured by the coreness profile and also by

the ENQ index as introduced in section 2.2. Geographical location can be proxied by the discrete and continuous measures while performance is reflected by publication, citation counts and the ratio of these two measures.

	Model(1)	Model(2)	Model(3)	Model(4)	Model(5)	Model(6)
NETPOS	*Coreness*	*Coreness*	*Coreness*	*Coreness*	*Coreness*	*Coreness*
GEOLOC	*Discrete*	*Discrete*	*Discrete*	*Continuous*	*Continuous*	*Continuous*
PERF	*Publication*	*Cit/Pub*	*Both*	*Publication*	*Cit/Pub*	*Both*
NETPOS(-1)	0,7548***	0,6547***	0,6556***	0,6556***	0,6548***	0,6556***
	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
Constant	−0,0008***	0,0038***	0,0020*	0,0017	0,0037***	0,0019*
	(0,0000)	(0,0005)	(0,0623)	(0,1041)	(0,0005)	(0,0730)
CONLOC				0,0855***	0,1674***	0,0852***
				(0,0000)	(0,0000)	(0,0000)
sq_CONTLOC				−0,0719***	−0,1498***	−0,0719***
				(0,0000)	(0,0000)	(0,0000)
DISCLOC_none	−0,0320***	−0,0143	−0,0036			
	(0,0000)	(0,2488)	(0,7706)			
DISCLOC_double	0,0253***	0,0433***	0,0207***			
	(0,0000)	(0,0000)	(0,0000)			
DISCLOC_center	0,0098***	0,0126**	0,0096*			
	(0,0075)	(0,0132)	(0,0558)			
PUB	0,0147***		0,0145***	0,0146***		0,0146***
	(0,0000)		(0,0000)	(0,0000)		(0,0000)
CIT_PUB		0,0001***	0,0001***		0,0001***	0,0001***
		(0,0008)	(0,0000)		(0,0009)	(0,0000)
CARR_Y		0,0057***	0,0047***	0,0050***	0,0058***	0,0047***
		(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
sq_CARR_Y		−0,0002***	−0,0002***	−0,0002***	−0,0002***	−0,0002***
		(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
N	108947	81603	81603	81301	81301	81603
Sum sq. residuals	5162.228	4039,123	3945.179	3932,128	4027.078	3945.179
S.E. of regression	0,2177	0,2225	0.2199	0,2199	0,2226	0.2199

1. Table – Regression results for coreness profile

Table 1 contains regression results when the coreness profile is used as the $NWPOS_{i,t}$ variable in EQ1. The model settings are summarized in the shaded area. We run regressions using the continuous and the discrete location measures. In case of using the discrete location measure, we employed a dummy variable approach to enter these discrete categories into the model. $DISCLOC\_double_{i,t}$ is a dummy variable which takes the value of 1 if author i in period t has double affiliation. $DISCLOC\_center_{i,t}$ takes the value of 1 if author i in period t is affiliated only in the center, while $DISCLOC\_none_{i,t}$ indicates if there is no information on the geographical position of author t in period i. Using this setting our reference group are those authors who are affiliated only in the geographical periphery. In case of using the continuous location measure we also include the squared version of the variable in order to see whether there is an inversed U-shaped relationship between location and network position indicating that belonging exclusively to the geographical

center or periphery is less favorable compared to the double affiliation. We use publication counts and the citation over publication measure of efficiency as proxies for scientific performance. As there is no significant correlation between the two performance measures we include them together as well as separately into the regression. Finally, we control for the age of authors by including the career year as a further explanatory variable ($CARR\_Y_{i,t}$) as well as the squared form in order to account for the higher performance of mid-career authors. Table 1 contains all the possible combinations of these settings.

The first thing to note in the regression results is that the lagged values of network position are always significant and positive, indicating a string persistence in network position over time. Second, publications and the efficiency of publication is always positive and significant, so it is reinforced by these results that a better performance contributes to network position: more publications and more citations per publication lead to a better connected network in terms of the coreness profile. It is interesting to add that the citations per publication measure is an ex post variable in the sense that it reflects the efficiency of the authors' publications in a given year as measured at the time of data retrieval – this means that although the number of citations on a publication takes time to increase to high levels, this kind of 'quality' is reflected almost immediately after publication by an improved network position. Third, we find a negative significant coefficient for the squared career year variable which indeed indicates a hump-shaped relationship between network position and age as evidenced by Figures 1 and 2.

When geographical location is proxied by the discrete measure (dummy variables, Models 1, 2 and 3), we see a positive significant coefficient for both center and double affiliated authors which means that both groups of authors have better network position in terms of coreness. However, the magnitude of the coefficient for double affiliated authors is much higher in all specifications which indicates that although authors who are affiliated in the geographical center have better network position than those in the geographical periphery, double affiliated authors (with affiliation both in the geographical center and periphery) even outperform exclusively center authors. This result reinforces the visual impressions obtained in the descriptive analysis. The regressions including the continuous measure for geographical location (models 4, 5 and 6) also reflect the same thing: the negative significant squared term indicates an inversed U-shaped relationship between location and network position. According to this those authors can improve their network position who have affiliations both in the geographical center and periphery. Authors who are affiliated exclusively in the geographical center or periphery have less favorable network positions.

	Model(7)	Model(8)	Model(9)	Model(10)	Model(11)	Model(12)
NETPOS	*log(ENQ)*	*log(ENQ)*	*log(ENQ)*	*log(ENQ)*	*log(ENQ)*	*log(ENQ)*
GEOLOC	*Discrete*	*Discrete*	*Discrete*	*Continuous*	*Continuous*	*Continuous*
PERF	*Publication*	*Cit/Pub*	*Both*	*Publication*	*Cit/Pub*	*Both*
NETPOS(-1)	0,7293***	0,6236***	0,6351***	0,6371***	0,6231***	0,6350***
	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
Constant	−0,0761***	−0,0144**	−0,0236***	−0,0253***	−0,0132**	−0,0228***
	(0,0000)	(0,0206)	(0,0001)	(0,0000)	(0,0333)	(0,0002)
CONLOC				0,4015***	0,7588***	0,4006***
				(0,0000)	(0,0000)	(0,0000)
sq_CONTLOC				−0,3435***	−0,6847***	−0,3449***

				(0,0000)	(0,0000)	(0,0000)
DISCLOC_none	−0,2146***	−0,0609	0,0010			
	(0,0000)	(0,3931)	(0,989)			
DISCLOC_double	0,0896***	0,2006***	0,1019***			
	(0,0000)	(0,0000)	(0,0000)			
DISCLOC_center	0,0352*	0,0691***	0,0561**			
	(0,0644)	(0,0032)	(0,0162)			
PUB	0,0628***		0,0632***	0,0632***		0,0634***
	(0,0000)		(0,0000)	(0,0000)		(0,0000)
CIT_PUB		0,0004***	0,0004***		0,0004***	0,0004***
		(0,0000)	(0,0000)		(0,0000)	(0,0000)
CARRY	0,0929***	0,0623***	0,0551***	0,0567***	0,0617***	0,0543***
	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
sq_CARRY	0,0006***	0,0007***	0,0008***	0,0008***	0,0007***	0,0008***
	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)	(0,0000)
N	88267	71187	71187	70956	70956	70956
Sum sq. residuals	82732.49	64737.64	63684.72	63602.63	64465.61	63405.22
S.E. of regression	0,9682	0,9537	0,9459	0,9468	0,9532	0,9454

2. Table -- Regression results for ENQ

Table 2 presents regression results when network position is measured by the ENQ index as presented in section 2.2. The results reinforce the impressions discussed previously. Network position measured by the ENQ index is also persistent, publication and efficiency have a positive impact on network position and there is a significant positive role for 'intermediate' geographical positions i.e. when an author is not exclusively affiliated in the geographical center or periphery – this latter effect is present irrespective of the proxy for geographical location. The only difference is that the coefficient of squared career years in now positive significant which indicates a positive (and nonlinear) trend in ENQ over the authors' lifetime which also evidenced by Figures 1 and 2.

The results presented in Table 1 and 2 seem robust as the magnitude of the coefficients does not change considerably across the different model specifications and using different proxies for the concepts result in similar evidences.

5. Summary

Text here

References

González-Pereira, B., Guerrero-Bote, V.P., Moya-Anegón, F. (2010): A new approach to the metric of journals' scientific prestige: The SJR indicator, *Journal of Informetrics*, 4(3), pp. 379-391.

Della Rossa, F., F. Dercole, F., Piccardi, C. (2013): Profiling core-periphery network structure by random walkers. *Scientific Reports*, 3, 1467, 2013.

Sebestyén, T., Varga, A. (2013a): Research productivity and the quality of interregional knowledge networks. *Annals of Regional Science*, 50(1), pp. 155-189.

Sebestyén, T., Varga, A. (2013b) A novel comprehensive index of network position and node characteristics in knowledge networks: Ego Network Quality. In Scherngell, Thomas (Ed.) *The geography of networks and R&D collaborations*. Springer, New York, 71-97.

Acedo, F.J., Barroso, C., Casanueva, C., Galán, J.L. (2006): Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. *Journal of Management Studies*, 43(5), pp. 957-983, 07.

Ahn, J., Oh, D., Lee, J. (2014): The scientific impact and partner selection in collaborative research at Korean universities. *Scientometrics*, 100(1), pp. 173-188.

Jeong, S., Choi, J.Y., Kim, J. (2011): The determinants of research collaboration modes: exploring the effects of research and researcher characteristics on co-authorship. *Scientometrics*, 89, pp. 967–983.

Fafchamps, M., van der Leij, M. J.,Goyal, S. (2010): Matching and Network Effects. *Journal of the European Economic Association*, 8, pp. 203–231.

Cunningham, S. W. Werker, C. (2012): Proximity and collaboration in European nanotechnology. *Papers in Regional Science*, 91, pp. 723–742.

Hardeman, S., Frenken, K., Nomaler, Ö., ter Wal, A. (2012): A proximity approach to territorial science systems. Paper prepared for the EUROLIO Conference on “Geography of Innovation” Saint-Etienne, France, 24-26 January 2012

Abbasi, A., Altmann, J., Hossain, L. (2011): Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5, pp. 594–607.

Yan, E., Ding, Y. (2009): Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), pp. 2107-2118.

Rodriguez, M.A., Pepe, A. (2008): On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Infometrics*, 2(3), pp. 195-201.

Abbasi, A., Chung, K.S.K., Hossain, L. (2012): Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, (48)4, pp. 671-679.