NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Nataliya N. Matveeva, Oleg V. Poldin

HOW NETWORK CHARACTERISTICS OF RESEARCHERS RELATE TO THEIR CITATION INDICATORS – A CO-AUTHORSHIP NETWORK ANALYSIS BASED ON GOOGLE SCHOLAR

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: EDUCATION
WP BRP 44/EDU/2017

Nataliya N. Matveeva[1], *Oleg V. Poldin*[2]

HOW NETWORK CHARACTERISTICS OF RESEARCHERS RELATE TO THEIR CITATION INDICATORS – A CO-AUTHORSHIP NETWORK ANALYSIS BASED ON GOOGLE SCHOLAR

The most common quantitative estimates of scientific performance are based on citation indices, and it is meaningful to identify what affects these indicators. In this work, we analyze the correlations between the citation characteristics of researchers and their co-authorship network parameters, which indicate the position of scientists in an academic network. To surpass the shortcoming of previous works we use a large sample and separate researchers by the year of their first citation. For constructing a co-authorship network, we used data about researchers from different disciplines, who have profiles in Google Scholar. The results of a count data regression model indicate that citations positively correlate with the number of co-authors, with position of the researcher in the co-authorship network (closeness centrality), and with the average number of co-author' citation. Also we reveal that the *h*-index and the i10-index are significantly associated with the number of co-authors and the average number of co-author citations. Based on these results, we can conclude that researchers who maintain more contacts and are more active than others have better bibliometric indicators on the average.

JEL Classification: A140; D830; Z130.

Keywords:co-authorship network; bibliometric analysis; Google Scholar; count data models

[1] National Research University Higher School of Economics. The Center for Institutional Studies; E-mail: nmatveeva@hse.ru

[2] National Research University Higher School of Economics. The Center for Institutional Studies

Introduction

In the process of research, scientists exchange ideas, discuss results and problems, and generate new concepts and theories. The results of such communication are often not tangible. What impact does such scientific collaboration have and how can this can be estimated? These questions are crucial for the investigation of scientific collaboration (Persson et al., 2004). Collaboration requires scientists to invest time and financial resources, but these expenses can be repaid by an increase in scientific output. A large number of conferences are held with the aim not only to exchange information, but also to intensify the collaboration between scientific groups.

Over the long-term, the importance of scientific results can be assessed, for example, by evaluating the impact on the economic development of society and the improvement of living standards. In the short-term the evaluation of scientific activity is measured by the relevance of publications. A widespread indicator of publications' relevance are citation indices and their derivatives (Hirsch, 2005; Egghe, 2006; Bergstrom et al., 2008; Bornmann, 2008; West et al., 2010). Kreiman and Maunsell propose criteria for assessing scientific output, which includes both quantitative and qualitative indicators based on reliable scientometric data and taking into account career stage (Kreiman&Maunsell, 2011). Fooladi et al. (2013) explored the influence of the journal impact factor on the quality of scientific articles. Scientific fields have different publication cultures and different approaches to cooperation so the evaluation of scientific activity is challenging.

Co-authorship networks are a form of social network reflecting collaboration between authors. In this network, authors are nodes and their joint publications are links. Scientific collaboration is easier to measure through co-authorship, because the product of co-authorship is a publication. Co-authorship networks allow not only visualize a interaction between the scientific community, but also to determine development trends, key participants and ways of spreading information. Joint publications represent the frequency of collaboration between researchers, and authors with the same number of publications often collaborate (Kretschmer, 1994). Glanzel and colleagues (Glanzel&Schubert, 2004; Glanzel, 2002) and Heffner (1981) investigated how network characteristics vary depending on the country and scientific field. Network construction and the methods for network evaluation have been described by Newman and colleagues (Newman, 2001; Girvan & Newman, 2002; Newman, 2004).

Collecting information about publications and co-authorship has become easier with the emergence of the scientific search systems and electronic databases of scientific citations such as Web of Science (WoS), Scopus, Microsoft Academic Search, Google Scholar. To date, there are a

number of studies dedicated to scientific databases (Meho&Yang, 2006; Kulkarni et al., 2009; Abbasi et al., 2011; Li et al., 2013; Uddin et al., 2013; Orduña-malea et al., 2015). The data in these databases differ in the number of indexed publications, coverage of scientific areas, methods of analysis, bibliometric characteristics and accessibility (Jacso, 2005; Bakkalbasi et al., 2006; Yang&Meho, 2006; Kulkarni et al., 2009; Bensman, 2013; Orduña-malea et al., 2015; Ortega, 2015).

In this study the electronic database Google Scholar (GS) was used. GS is a reliable tool to analyze highly–cited publications (Martin-Martin et al., 2017) and more appropriate for bibliometric analysis of young researchers because it indexes more type of publications (Franceschet, 2009). In contrast to WoS and Scopus this database is open and contains elements of social networks (authors can manage their profile and manually indicate the publications and co-authors) and it indexes more sources although not all sources have passed peer review (proceedings, preprints, thesis) and some cannot be related to scientific publications (popular science notes and blog records). Despite such issues, bibliometric indicators calculated based on GS are comparable with the similar indicators for WoS (Wildgaard. 2015; Harzing&Van der Wal, 2008) Possible reason for this the citation rates of non-peer reviewed articles are much lower than for peer reviewed articles.

To date, there are very few studies considering co-authorship networks and scientific citation that are based on GS profiles (Arnaboldi et al., 2016; Hossain&Kobourov, 2017). Possible reasons for this are the misidentification of authors and their works, the presence of phantom authors, and the incorrect identification of the year of publication, which affects the citation indexes (Jacso, 2008a). In the last decade, GS has made great efforts to make their database more accurate https://scholar.google.com/intl/en/scholar/citations.html).

Information and knowledge spread through co-authorship networks representing the social resources of scientists. It is reasonable to assume that access to resources facilitates scientific output. Studies have shown that there is a positive correlation between some network characteristics (normalized degree centrality, normalized eigenvector centrality, average tie strength) and citation index (Abbasi et al., 2011), between the citation count of article and tie strength between authors (Uddin et al., 2013). Li (2013) and Guan et al. (2017) found that research collaboration positively affect paper's citation count.

Here we evaluate the correlation between the network characteristics of the researchers such as degree centrality, closeness centrality and citation indicators (the total number of citations, the *h*-index, the i10-index). In addition, we identified the correlation between citation indicators and the

average number of co-author citations. The network parameters indicate the position of the researchers in the network of collaboration and characterize the potential of the researchers. For example, the investigation of the research team and their efficiency (citation indicators) can indicate the potential career development of young scientists. The main question of this study is how the network parameters of relatively young researchers relate to their citation indicators. For this purpose, we divided the researchers by scientific field and by the year of their first citation. This makes the sample more homogeneous in terms of citation tenure and scientific field suppressing interfering factors within subsets. Citation tenure is calculated as the number of years in which a researcher's publications were cited.

The drawbacks of some earlier studies include using a small sample of scientists from different fields (Zuckerman, 1967) or using data about scientists from only one scientific field (Avkiran, 2013; Uddin et al., 2013; Puuska et al., 2014; Yu et al., 2014); these affect the reliability of the results. The co-author network constructed on the basis of parsed articles also affect the quality of the analysis due to complications in the determination of the conformity between authors with identical surnames and their works (Méndez-Vásquez et al., 2012).

To fill the void in the literature, we use a richer dataset (110,000 for network analysis), taking into account the size of sample (34,000 for regression analysis). The sample was divided into subsets corresponding the citation tenure and scientific field because citation indicators depend on these factors (Ductor et al., 2014; Murugesan et al., 1978). Moreover, we investigate not only how scientific collaboration is linked to citation indices but we also take into account a new parameter – the average number of co-author citations and investigate how this correlates with the citation count for relatively young researchers.

Study context and data

Google Scholar

We constructed a co-authorship network based on the profiles of researchers in GS, a free web search system, which provides full-text search of all types of publications from various disciplines. This system was launched in November 2004. GS contains basic information about publications such as title, authors and total number of citations. Information about citing sources, publisher, year of publications and links on full-text publication is also provided by GS.

GS comprises features of an online social networking site, since it allows users to register their profiles, in which they can define their co-authors in a special section of the personal page.

The personal page of researchers contains data about publications indexed by GS, their citations, i10-index, and *h*-index for the all years and for the last five years, and optional information about researcher affiliation, position and scientific interests. Registered researchers can manage the list of their publications and the list of co-authors.

As a tool for bibliometric analysis GS has several advantages in comparison with WoS and Scopus. First, GS indexes more scientific sources including PhD-thesis, Master-thesis, arXiv and other preprints in various languages. Secondly, GS is updated the data more often. Thirdly, it has free access and allows the extraction of information by creating web parser programs.

The drawbacks of GS are the indexing of non-scientific sources (blogs and presentations), the indexing of the articles in non-refereed journals and proceedings, some profiles have misidentified authors and their works. These lead to incorrect citation counts (Jacso, 2008a; Jacso, 2008b). It can be assumed that these drawbacks are not prevalent since there is a strong correlation between bibliometric indicators based on GS and based on WoS (Harzing&Van der Wal, 2008; Wildgaard, 2015).

There are some limitations in the structure of GS profiles: the fields of position, affiliation, scientific interests and list of co-authors are not required; the misidentification of scientists last name and their works, no career or affiliation tracking. The latter two problems are also relevant to WoS and Scopus. Despite these limitation GS profiles automatically and quite accurately display bibliometric indicators of scientists. For profiles with an empty list of co-authors, it is possible to restore this list using other profiles of co-authorship network.

Citations and their derivatives

Citations of a publication are references to this publication in other papers. The scientist's citation index is sum of references of all their publications. In contrast to the number of publications, citations take into account the qualitative aspect of scientific activity. It has been shown that better works are cited more often (Garfield, 1979). The major shortcoming of citations is the dependence of this indicator on the disciplines and on the year of publication. The mention of someone's work can be in a critical manner, pointing to weaknesses of the work (Toutkoushian & Webber, 2011). The citation indices also vary according to which bibliographic database is used (Bakkalbasi et al., 2006).

The *h*-index is a popular quantitative characteristic of a scientist's productivity, since it takes into account both the number of publications and their citations. A researcher has an *h*-index of *h* if

they have *h* publications which have been cited at least *h* times each (Hirsch, 2005). The main drawback of this index is that does not reflect the citation number of highly cited publications (Costas & Bordons, 2007).

In addition to the *h*-index, GS uses the i10-index, which indicates the number of scientific publications which have at least 10 citations. It was introduced in 2011 and is calculated only by GS. It is a more accurate (in comparison with the *h*-index) reflection of the number of highly cited publications, and i-10 is easy to calculate. The shortcoming of i-10 index is its local use and, like the *h*-index, it does not reflect the publication having a very high number of citations.

GS indexes not only journal articles and English-language works, but also other types of publications, giving citation indicators higher than those in WoS and Scopus. However, there is a strong correlation between citation ratings compiled by these databases (Franceschet, 2009; Delgado & Repiso, 2013; Wildgaard, 2015).

Co-authorship network

In a co-authorship network, nodes represent the authors (letters A-E in Figure 1) and their joint publications are links (1-5 links in Figure 1). Such a network represents the interaction between authors: how often authors have collaborated (represented by publications), who is the most active, the position of authors in the network, how information can be distributed and so on. Parameters of the network characterize the quantitative estimate of these interactions.

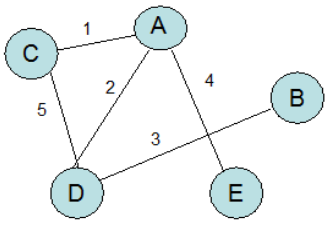

Fig. 1: Co-authorship network illustration

This study used information about co-authors and scientific interest, indicated in their profiles in GS for building the network. Based on scientific interest, scientists were divided into groups corresponding to the fields of computer science, economics &finance, biology &medicine, physics &chemistry, mathematics, and social sciences &humanities.

The total number of nodes, degree centrality, average number of co-authors citation, and closeness centrality were calculated for the full co-authorship network and for each discipline. The

total number of nodes is the number of scientists who have a publication. This number is different in various disciplines and reflects the presence various disciplines in GS. Degree centrality is the number of co-authors a scientist has. The degree centrality of nodes is measured as the sum of its adjacent nodes (Scott, 1991). The average number of co-author citations indicates how many citations the co-authors have on average in the sample. Closeness centrality measures how close a particular author is to another.

The data collection started with a list of 12 profiles of Russian researchers, and then information was extracted about their co-authors, co-authors of co-authors and so on up to 110,000. Data collection was carried out from February to March 2015 using a web-parser written on MATLAB. Based on these data an adjacency matrix was made. In this matrix rows and columns are nodes and elements of the matrix indicating whether the pairs of vertices are connected or not. After preparing the adjacency matrix, we calculate short paths for each node using Dijkstra's algorithm (Dijkstra, 1959). In addition, we calculated network parameters and citation indicators.

An estimation of empirical models was made for 34,000 profiles taking into account the links with all 110,000 network participants. For the regression analysis we use researchers whose year of the first citation is detected (34,000 out of 110,000) because we estimate citation tenure as response variable for citation indicators. In GS, information about a researcher's first citation can be extracted only for authors with the first citation after 2007.

Analysis and Results

Using our data, we calculate network metrics and citation indicators. The results for the full sample and for groups of disciplines are shown in Table 1. The highest average values of citations, for the *h*- and i10- indices are in the fields of biomedicine, physics &chemistry, and the lowest indices are in economics &finance. The average number of co-authors for all disciplines is 6, this indicator is above average in the computer science & math and below average in economics & finance. For the full sample closeness centrality is 0.185, this is above average in computer science and below average in economics & finance and physics & chemistry. The researchers from computer science prevail in our sample. So, the result obtained by Ortega (2015) was confirmed in that researchers who are engaged in the field of computer sciences are the most in GS.

Tab. 1. Descriptive statistics: mean (std. deviation)

	Full sample	Disciplines					
		BM	CS	EF	Math	PhCh	SH
Citation	333.9 (1041.9)	542.8 (1529.2)	262.1 (873.4)	234.1 (414.0)	266.0 (504.3)	470.3 (1153.3)	307.6 (611.7)
h-index	6.4 (4.94)	7.6 (5.6)	5.7 (4.3)	5.7 (3.8)	6.5 (4.3)	7.9 (5.8)	6.542 (4.755)
i10 - index	5.8 (9.3)	7.8 (10.3)	4.8 (7.3)	4.7 (5.7)	5.9 (7.2)	8.4 (11.9)	6.113 (8.597)
Number of co-authors	6.03 (6.62)	5.93 (6.34)	6.65 (7.59)	4.07 (3.75)	6.11 (5.67)	5.93 (5.39)	5.913 (6.521)
Average number of co-authors citation	4285.2 (5728.4)	6355.0 (7630.2)	3705.9 (4498.8)	4209.0 (6239.2)	3984.9 (4600.6)	5060.1 (5720.8)	4019.3 (5194.9)
Tenure	4.928 (1.725)	5.111 (1.651)	4.758 (1.746)	5.221 (1.672)	5.088 (1.669)	5.041 (1.661)	5.081 (1.671)
Ln(tenure)	1.508 (0.466)	1.557 (0.431)	1.466 (0.482)	1.573 (0.454)	1.550 (0.439)	1.542 (0.429)	0.180 (0.0186)
Closeness centrality	0.185 (0.0163)	0.185 (0.0141)	0.190 (0.0136)	0.165 (0.0188)	0.185 (0.0154)	0.178 (0.0134)	0.180 (0.0186)
Observations	34701	1878	6514	1004	1247	1353	2259

BM — Biology and Medicine, CS — Computer Science, EF — Economics and Finance, Math — Mathematics, PhCh — Physics and Chemistry, SH — Social Sciences and Humanities

Unlike the citation parameter, the number of co-authors or the centralities of authors depend significantly on which part of the co-author's network is outside the analyzed data. In the present study, data on the connected co-authorship network were used, which allowed us to minimize the bias of network characteristic estimates.

An important variable that affects the citation indices is the author's citation tenure. The presence in sample of authors with various citation tenures affects their bibliometric indicators (Ductor et al., 2014). The logarithm function allows us to minimize the variance of this parameter. For our analysis, we separate the sample by the year of first citation to minimize time effect of citation indicators.

Empirical models

The dependent variables represent the number of citations and therefore are positive integers. For regression analysis count data models are often used. We use model variations with negative binomial distribution, which is a generalization of the simpler and limited Poisson model (Cameron & Trivedi. 2013). In particular, the model with negative binomial distribution has more flexibility for describing the variance of the dependent variable and is more appropriate for modeling situations in which individual citation indicators are interrelated events (Ajiferuke & Famoye, 2015). For the number of citations as dependent variables we ran the Zero-truncated model because we do not take into account zero citations. Not all scientists in the sample have publications with 10 or more citation and for that reason for i-10 index the Hurdle model were used.

Negative binomial model. The probability function of a variable with a negative binomial distribution can be written as:

(1)

where Γ is the gamma function.

In the regression model, it is assumed that mean and variance of the dependent variable depend on the explanatory variables through the exponential function:

(2)

Zero-truncated model. We do not consider authors with zero citations so when the dependent variable is the citation of an author, the sample contains only positive values. The conditional probability distribution function with excluded zero values of the dependent variable is:

(3)

where *f(y)* is a probability function with zero values included. This adjustment is taken into account in zero-truncated models.

Hurdle model. The i10-index estimates the number of publications which have 10 or more citations. Since not all authors have such publications, for the empirical modeling of this index Hurdle models were used. These models consist of two components. The first component describes

the distribution of the dependent variable with zero values and the second the non-zero distribution of the dependent variable. The probability density function is given by:

$$f(y) = \begin{cases} f_1(y), & if \;\; y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y), & if \;\; y \geq 1 \end{cases} \tag{4}$$

In the regression models, we estimate two equations. The first equation represents a logistic model of a binary choice for the probability of the author having a non-zero value for the i10-index. The second component describes distribution of non-zero values using a negative binomial model with truncated zeros.

2. Results

For the correlation analysis of the citation index, we used a model with negative binomial distribution and zero-truncated citation values. The results for three model specifications of the full sample are shown in Table 2. The model specifications differ in a number of response variables. In specification 1, the characteristic of author's position in co-authorship network is the number of his direct co-authors, in specification 2 the closeness centrality (see formula (1)), in specification 3 both network parameters. The common variables in all specifications are the average number of citations of co-authors and the logarithm of author's citation tenure. The results reveal that closeness centrality has a minor role relative to the number of co-authors. In the full specification (column 3 of Table 2), closeness centrality is not significant, but it becomes significant if the number of co-authors is excluded from the model. In this case, closeness centrality is a single variable which describes the position of authors in network. The estimates of the specification without closeness centrality (column 1) and full specification (column 3) are similar. The average number of citations of co-authors, and citation tenure are significant at 1% level.

Tab. 2. Regression estimates for citation index

	(1)	(2)	(3)
Average number of co-authors citation $\times 10^{-4}$	0.854***	0.817***	0.862***
	(0.033)	(0.035)	(0.035)
Number of co-authors	0.061***		0.063***
	(0.002)		(0.003)
Closeness centrality		14.84***	-1.066
		(0.726)	(1.115)
Ln(tenure)	2.109***	2.247***	2.108***
	(0.040)	(0.036)	(0.040)
Constant	1.307***	-1.212***	1.491***
	(0.0634)	(0.153)	(0.213)
Ln(α)	0.160***	0.240***	0.160***
	(0.021)	(0.018)	(0.022)
N	34701	34701	34701
pseudo R^2	0.0543	0.0482	0.0543

In brackets are standard errors; *** $p< 0.01$

For the interpretation of results, it is more convenient to consider not the variable coefficients in the exponential function in model (4), but changes of the dependent variable with increments of the explanatory factors. Since response variables differ in scale, we considered the effects of a discrete increment of these variables from their mean values by one standard deviation, and other variables are fixed at the mean value.

The mean of the effects and the levels of 95% confidence intervals for the full sample and for individual specializations are shown in Table 3. In the full sample, an increase in the average number of co-author citations to a standard deviation of 5,728 is associated with an increase of 116.8 citations, with an increase in the number of co-authors from 6 to 13, citations increases by 92.4. In our specializations, the greatest absolute effect is observed for authors working in the field of biology &medicine, the lowest in the field of economics &finance. However, the mean and variance of citations for biology &medicine are greater than for economics &finance. In the last column of Table 3, the absolute effects are normalized to the standard deviations of the citations in the corresponding sample. The figures show how many standard deviations the citation changes by with the growth of the response variable by one standard deviation. In these terms, the average number of co-author citations has the greatest effect for specialization in mathematics and the number of co-authors has the greatest effect for economics &finance. Quantitative estimates of the

number of co-authors vary from 0.09 to 0.17, the coefficients of co-authors citations vary from 0.06 to 0.17.

Tab. 3. Changes of citation with increasing response variables by 1 std.

	Δy	Low level	Higher level	Δy/σ
Full sample				
Average number of co-authors citation	116.8	103.9	129.7	0.11
Number of co-authors	92.4	86.1	98.8	0.09
Ln(tenure)	309.9	292.7	327.0	0.30
Computer science				
Average number of co-authors citation	65.5	49.1	81.9	0.07
Number of co-authors	105.8	91.6	120.1	0.12
Ln(tenure)	223.4	206.1	240.6	0.26
Economics and Finance				
Average number of co-authors citation	24.7	6.5	42.8	0.06
Number of co-authors	71.6	55.3	88.0	0.17
Ln(tenure)	266.7	231.0	302.4	0.64
Biology and Medicine				
Average number of co-authors citation	176.2	123.5	228.8	0.12
Number of co-authors	156.5	117.4	195.5	0.10
Ln(tenure)	483.4	416.6	550.2	0.32
Physics and Chemistry				
Average number of co-authors citation	124.4	85.6	163.2	0.11
Number of co-authors	112.0	81.5	142.5	0.10
Ln(tenure)	420.8	355.3	486.4	0.36
Math				
Average number of co-authors citation	86.8	44.6	129.1	0.17
Number of co-authors	80.1	61.5	98.7	0.16
Ln(tenure)	260.4	218.9	302.0	0.52

Social Sciences and Humanities				
Average number of co-authors citation	93.4	46.5	140.2	0.15
Number of co-authors	88.2	69.2	107.2	0.14
Ln(tenure)	223.1	137.1	309.0	0.36

We ran a negative binomial regression model taking into account zero values to identify which of the response variables (average number of co-authors citation, number of co-authors, and logarithm of citation tenure) impact the *h*-index. The results of this model for the full sample are present in Table 4. All the variables are significant. The negative binomial regression coefficients are interpreted as follows: the number of co-authors has a greater affect on the *h*-index than co-author citations, while for the citation (Table 3) the effects of these variables are comparable. Generally, the normalized effect of the citation tenure is large (0.66) and twice as much as for citations (0.30).

Tab. 4. Regression estimates for h- index

	Coefficient β	Δy if Δx change by 1 std			
		Δ*y*	Low level	Higher level	Δ*y*/*σ*
Average number of co-authors citation $\times 10^{-4}$	0.173^{***} (0.010)	0.58	0.52	0.65	0.12
Number of co-authors	0.031^{***} (0.001)	1.25	1.18	1.32	0.25
Ln(tenure)	0.985^{***} (0.008)	3.26	3.20	3.32	0.66
Constant	-0.021^{*} (0.012)				
$\ln(\alpha)$	-2.246^{***} (0.035)				
Number of observations	34701				
pseudo-R^2	0.123				

In brackets are standard errors; * p < 0.1,*** p < 0.01

The results of Hurdle models for the i10-index are shown in Table 5. The first part is the logistic model of a binary choice in which the probability to has the author a non-zero i10-index (ie the author has at least one publication with 10 citation) is the dependent variable. The correlation coefficient values are interpreted as follows: with an increase in the average number of co-author citations by one standard deviation the probability of having a non-zero index increases by 2.6%,

and by 6% with an increase of the number of co-authors by one standard deviation. In the second part, we run a negative binomial model with truncated zeros in which a non-zero value of the i10-index is the dependent variable. The last column show that for the i10-index the normalized effect of a discrete change in the average number of co-author citations (0.12) is noticeably weaker than a change in the number of co-authors (0.22).

Tab. 5. Regression estimates for i-10 index

	Coefficient β	Δy if Δx change by 1 std			
		Δy	Low level	Higher level	$\Delta y/\sigma$
Dependent variable – P(i10>0)					
Average number of co-authors citation $\times 10^{-4}$	0.730^{***} (0.060)	0.026	0.023	0.030	
Number of co-authors	0.208^{***} (0.008)	0.060	0.058	0.062	
Ln(tenure)	2.804^{***} (0.040)	0.058	0.056	0.061	
Constant	-3.374^{***} (0.062)				
Number of observations	34701				
pseudo-R^2	0.342				
Dependent variable – i10 | i10>0					
Average number of co-authors citation $\times 10^{-4}$	0.365^{***} (0.019)	1.156	1.021	1.292	0.12
Number of co-authors	0.050^{***} (0.001)	2.024	1.921	2.128	0.22
Ln(tenure)	1.861^{***} (0.030)	4.093	3.925	4.262	0.44
Constant	-1.909^{***} (0.051)				
$\ln(\alpha)$	-0.384^{***} (0.030)				
Number of observations	28690				
pseudo-R^2	0.0738				

In brackets are standard errors; *** $p < 0.01$

For all the citation indices (the number of citations, the *h*-index and the i10-index) there is a significant correlation between the number of co-authors and the average number of co-author citations. For the citation indicator, the results show that closeness centrality has a minor influence on the number of co-authors, and the average co-author citations has the greatest affect on the citation indicator for specialization in mathematics; for the number of co-authors in economics

&finance. For the *h*-index, the number of co-authors has a greater affect (0.031) on the *h*-index than average co-author citations (0.173). For the i10-index, the average co-author citations has a noticeably weaker affect on the i10-index than the number of co-authors. The position of a researcher in the co-authorship network and co-author citations correlate to their scientific impact expressed in citation indicators.

Conclusion

In this work, we investigated how scientific collaborations represented by a co-authorship network relate to scientific impact. For this purpose, we analyzed correlation between the number of co-authors, the author's centrality, the average number of co-author citations, citation tenure and the number of citations measured by the *h*-index and the i10-index. To avoid the limitations of previous research caused by limited datasets we used a large dataset containing researchers from various countries and scientific fields. We divided researchers by the year of first citation and ran a regression analysis for relatively young scientists.

The results of the count data regression model show that there is a positive correlation between researcher citation counts and the number of co-authors, between citations and the author's centrality. It was also found that the average number of co-author citations has a positive effect on a researcher's citations. We also estimated the extension of the statistical connection between the variables by taking the normalized coefficients that characterize the variation of the dependent variable (in absolute figures and standard deviations) with an increase in the factors by one standard deviation. The variations in results depend on the scientific specialization of the author. In addition, in this study the *h*-index and the i10-index are used as indicators of scientific impact. These indicators significantly correlate with the number of co-authors and average co-author citations, while the normalized effect of the number of co-authors is approximately twice the effect of the average number of co-author citations.

We conclude that scientists who maintain more contacts and are more active than others have better bibliometric indicators on average. These results correspond to other studies, for example (Glanzel, 2002, Persson et al., 2004), which show that citations grow with an increase in the number of co-authors and more frequently cited papers are mostly co-authored by scientists who have higher network characteristics (Uddin, 2013; Cimenler, et al., 2014).

On the other hand, our results do not coincide with Abbasi et al. (2011) who concluded that closeness centrality is not positively correlated with citation indicators. A possible reason for that is the inclusion in their sample of researchers with different citation tenures. For example the presence

in co-authorship network of professors with a large number of co-authors can influences the centrality of their graduate students. In our work we obtain the opposite results due to the separation of the sample by year of first citation.

The results allow to us formulate several hypotheses that require further evaluation. The correlation between the citations and the number of co-authors, and the average number of co-author citations can be explained by the influence of the number of co-authors (the more co-authors, the more citations the work gets) and the influence of the quality of the co-authors on the citation of the joint publication (the more citations an author has than more likely their work will be cited). On the other hand, when building a team of co-authors there is self-selection: complex scientific problems require an appropriate number and quality of researchers, so if researchers work with highly-cited co-authors it can be explained by the demand for their own abilities.

This research is not without its limitations. All the measured indicators are size-dependent (they all depend on the number of publications of an author). For future study the number of publications of each author should be controlled. Not all researchers have a GS account, which may affect the network because of missing nodes.

The revealed statistical correlations do not allow us to identify a cause and effect relationship between variables. A fundamental problem is in the endogeneity of social ties in general and in co-authorship in particular, because links are formed on a voluntary basis. On the dataset (parsed for one time point), it is impossible to obtain dynamic data about the variables. Nevertheless, based on the sample of relatively young scientists, which do not have higher citation indicators, we can suggest that, in a greater degree, network measures affect the citation index.

However, this work does not allow us to disclose the cause-effect relationships but it shows that the research team in which a scientist works and the citation of their work are interrelated. Thus, correlation analysis allows us to understand how the formation of links in network collaboration relates bibliometric characteristics.

This study is the basis for future investigations on the scientific impact of co-authorship and citation networks. In the future work we intend to collect dynamic data about scientists in GS and analyze the evolution of the co-authorship characteristics its scientific impact. Such studies are crucially important for understanding how scientific collaboration influences scientific impact.

References:

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 (4), 594–607.

Ajiferuke I., Famoye F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. Journal of Informetrics, 9 (3), 499–513.

Avkiran N. K. (2013). An empirical investigation of the influence of collaboration in finance on article impact. Scientometrics, 95 (3), 911–925.

Arnaboldi, V., Dunbar, R. I., Passarella, A., & Conti, M. (2016, January). Analysis of co-authorship ego networks. In International Conference and School on Network Science (pp. 82-96). Springer, Cham.

Bakkalbasi N., Bauer K., Glover J., Wang L. (2006). Three options for citation tracking: Google scholar, Scopus and Web of Science. Biomedical Digital Libraries, 3 (1), 7.

Bensman S. J. (2013). Eugene Garfield, Francis Narin, and Pagerank: The theoretical bases of the Google search engine. arXiv preprint arXiv:1312.3872.

Bergstrom C. T., West J. D., Wiseman M. A. (2008). The eigenfactor metrics. The Journal of Neuroscience, 28 (45), 11433–11434.

Bornmann L., Daniel H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64 (1), 45–80.

Cameron A. C., Trivedi P. K. (2013). Regression analysis of count data. Cambridge University Press.

Costas, R., &Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of informetrics, 1(3), 193-203.

Delgado E., Repiso R. (2013). The impact of scientific journals of communication: Comparing Google Scholar metrics, Web of Science and Scopus/el impacto de lasrevistas de comunicación: Comparando Google Scholar metrics, Web of Science y Scopus. Comunicar, 21 (41), 45–52.

Ductor L., Fafchamps M., Goyal S., van der Leij M. J. (2014). Social networks and research output. Review of Economics and Statistics, 96 (5), 936–948.

Egghe L. (2006). Theory and practice of the g-index. Scientometrics, 69 (1), 131–152.

Fooladi M., Salehi H., Yunus M. M., Farhadi M., Chadegani, A., Farhadi H., Ebrahim N. (2013). Does criticisms overcome the praises of journal impact factor? Asian Social Science, 9 (5), 176–182.

Franceschet M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics, 83 (1), 243–258.

Garfield E. (1979). Citation indexing: Its theory and application in science, technology, and humanities, New York: Wiley.

Cimenler, O., Reeves, K. A., &Skvoretz, J. (2014). A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. Journal of Informetrics, 8(3), 667-682.

Girvan M., Newman M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99 (12), 7821–7826.

Glänzel W. (2002). Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. Library trends, 50 (3), 461–475.

Glänzel W., Schubert A. (2004). Analysing scientific networks through co-authorship, Handbook of quantitative science and technology research. Springer, pp. 257–276.

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics, 11*(2), 407-422.

Harzing A.-W. K., Van der Wal R. (2008). Google Scholar as a new source for citation analysis. Ethics in science and environmental politics 8 (1), 61-73.

Heffner A. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. Scientometrics, 3(1), 5-12.

Hirsch J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America, 102 (46), 16569–16572.

Hossain, M. I., &Kobourov, S. (2017). Research Topics Map: rtopmap. arXiv preprint arXiv:1706.04979.

Jacso P. (2005). As we may search–comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Current Science, 89 (9), 1537–1547.

Jacsó P. (2008a). The pros and cons of computing the h-index using Google Scholar. Online information review 32 (3), 437–452.

Jacso P. (2008b). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for FW Lancaster. Library trends, 56 (4), 784–815.

Kreiman G., Maunsell J. H. R. (2011). Nine criteria for a measure of scientific output. Beyond open access: visions for open evaluation of scientific papers by post-publication peer review, 11.

Kretschmer H. (1994). Coauthorship networks of invisible colleges and institutionalized communities. Scientometrics, 30 (1), 363–369.

Kulkarni A. V., Aziz B., Shams I., Busse J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. Jama, 302 (10), 1092–1096.

Li E. Y., Liao C. H., Yen H. R. (2013). Co-authorship networks and research impact: A social capital perspective. Research Policy, 42 (9), 1515–1530.

Martin-Martin, A., Orduna-Malea, E., Harzing, A. W., &López-Cózar, E. D. (2017). Can we use Google Scholar to identify highly-cited documents?. Journal of Informetrics, 11(1), 152-163.

Meho L. I., Yang K. (2006). A new era in citation and bibliometric analyses : Web of science, Scopus , and Google Scholar. arXiv:cs/0612132.

Méndez-Vásquez R. I., Suñén-Pinyol E., Cervelló R., Camí J. (2012). Identification and bibliometric characterization of research groups in the cardio-cerebrovascular field, Spain 1996–2004. Revista Española de Cardiología (English Edition), 65 (7), 642–650.

Murugesan P., Moravcsik M. J. Variation of the nature of citation measures with journals and scientific specialties //Journal of the American Society for Information Science. – 1978. – Т. 29. – №. 3. – С. 141-147.).

Newman M. E. J. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98 (2), 404–409.

Newman M. E. J. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69 (6), 066133.

Orduña-malea E., Ayllón J. M., Martín-martín A., López-cózar E. D. (2015). Methods for estimating the size of Google Scholar. Scientometrics, 104 (3), 931–949.

Ortega J. L. (2015). How is an academic social site populated? A demographic study of Google Scholar citations population. Scientometrics, 104 (1), 1–18.

Persson O., Glänzel W., Danell R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60 (3), 421–432.

Puuska H.-M., Muhonen R., Leino Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. Scientometrics, 98 (2), 823–839.

Scott, J. 1991). Social network analysis: a handbook.: Sage.

Toutkoushian R. K., Webber K. (2011). Measuring the research performance of postsecondary institutions, in C. J. Shin, K. R. Toutkoushian, and U. Teichler, (eds.), University rankings: Theoretical basis, methodology and impacts on global higher education. Dordrecht: Springer Netherlands, pp. 123–144.

Uddin S., Hossain L., Rasmussen K. (2013). Network effects on scientific collaborations. PLoS ONE, 8 (2),

West J., Bergstrom T., Bergstrom C. T. (2010). Big Macs and eigenfactor scores: Don't let correlation coefficients fool you. Journal of the American Society for Information Science and Technology, 61 (9), 1800–1807.

Wildgaard L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in astronomy, environmental science, philosophy and public health in Web of Science and Google Scholar. Scientometrics, 104 (3), 873–906.

Yang K., Meho L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. Proceedings of the American Society for Information Science and Technology, 43 (1), 1–15.

Yu Q., Shao H., Long C., Duan Z. (2014). The relationship between research performance and international research collaboration in the C&C field. Experimental and Clinical Cardiology, 20 (6), 145–153.

Zuckerman H. (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. American Sociological Review, 32 (3), 391–403.

Contact details and disclaimer:

Nataliya N. Matveeva

National Research University Higher School of Economics (Moscow, Russia). The Center for Institutional Studies.
E-mail: nmatveeva@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.