Applied Network Science

RESEARCH Open Access

The role of highly intercited papers on scientific impact: the Mexican case

Rodrigo Dorantes-Gilardi[1], Aurora A. Ramírez-Álvarez[2] and Diana Terrazas-Santamaría[2*]

*Correspondence:
dterrazas@colmex.mx

[1] Department of Physics, Network Science Institute, Northeastern University, Boston, MA, USA
[2] Center for Economic Studies, El Colegio de México, Mexico City, Mexico

Abstract

The present paper explores the relationship between highly intercited papers in the *k*-max of citation networks and an author's impact from the Mexican National System of Researchers (SNI). We investigate whether a more interconnected network, a higher *k* of the *k*-max, explains the variation of the total number of citations, controlling for personal characteristics such as SNI level, area of expertise, and the number of publications. We find that the *k*-max is positively and significantly correlated with impact. In this context, we find that the share of self and collaborator-citations increases with the magnitude of the *k*-max and women tend to have less interlinked cores of their citation networks than men (smaller *k*'s). Interestingly, we find that women tend to have a higher share of third-party citations while men tend to have a higher share of self and collaborator-citations, for all *k*'s and areas of expertise. We conduct a Blinder–Oaxaca decomposition to better understand the citation gender gap and find that much of it can be explained through the differences in observable characteristics (including the *k*-max) between women and men.

Keywords: Gender inequality, Citation network, *k*-core, Academic impact

Introduction and motivation

Mexico frequently emerges as one of the most unequal countries across various economic, societal, and demographic topics (Campos-Vázquez et al. 2018). Science is not the exception where striking gender differences have been reported not only regarding women representation in general, but specially within the highest levels of the decision-making ladder (Lloyd 2018; Sandoval-Romero and Larivière 2020; Rodríguez Miramontes et al. 2017). In 2018, only 37% of the national researchers' system were women, their presence varying across fields and seniority (CONACYT 2018).

The number of citations in the evaluation and promotion of researchers has gained importance in recent decades in which different strategies to magnify these have been uncovered, differing significantly between genders (van den Besselaar and Sandström 2016; King et al. 2017; Fortunato et al. 2018; Huang et al. 2020). The academic career has been regarded as a publish-or-perish game where a scientific contribution is measured against the impact it creates.

In this context, Ioannidis et al. (2019) explain the need for citation metrics that are field-adjusted, accounting for extreme self-citations and able to detect *citation farms* in which a small cluster of authors disproportionately cites each other's papers. The strategic use of self-citations to boost an author's impact has been widely studied (Kacem et al. 2020; Van Noorden and Singh Chawla 2019; Wallace et al. 2012). For instance, King et al. (2017) find that men self-cite 56% more than women, creating an asymmetric cumulative advantage between genders.

Reciprocity between authors within their closest social circle has been less studied and only recently has gained importance. Li et al. (2019) investigate whether authors that present high citation reciprocity (exchange of citations between authors) outperform their peers and find that only those in the lowest part of the citation distribution do benefit from this strategy to boost their visibility.

Therefore, the issue of exploring the determinants of an author's impact is a mix of personal attributes and the authors' social network. It is well established that collaboration in academia is mostly beneficial for all parties since it can improve their productivity and impact through the acquisition of resources, learning of new abilities, boost of citations, and a more lengthy career (Van Der Wal et al. 2021; Wallace et al. 2012; Fortunato et al. 2018; Paraskevopoulos et al. 2021; Dorantes-Gilardi et al. 2021).

The study of real-world networks has stressed the necessity of developing centrality, rankings, and structural organization measures to uncover complex connectivity patterns usually hindered and are proven useful to characterize different network structures and configurations (Alvarez-Hamelin et al. 2005). One macro-level measure to find interconnected links within the network is based on its k-core, defined as the maximal set of nodes that have at least degree k within the set (Kong et al. 2019; Seidman 1983). The concept of k-core has proven helpful in a variety of financial, biological, and community detection topics (Kong et al. 2019; Giatsidis et al. 2011; Burleson-Lesser et al. 2020; Dorantes-Gilardi et al. 2021). In practice, it is of particular interest the study of the maximal degree k such that a k-core exists, the so-called main core and k-max, since the nodes belonging to it are responsible for providing a "structure" to the network due to their high relations strengths (Burleson-Lesser et al. 2020). We will only consider the k-max throughout the paper.

In this paper, we explore whether an author's k-max (magnitude of k) correlates with the number of citations, once we control for other individual characteristics, such as the number of papers, area of expertise, the average number of co-authors per paper, career length, among others. We construct personal citation networks where the nodes are papers and links are citations that may come from a paper of the same author (self-citation), a paper of a co-author (collaborator-citation), or someone else (third-party citation).

Our goal is to uncover whether a more interlinked inner core (higher k) implies more citations. To the best of our knowledge, this is the first paper to investigate the link of the citation networks topology through differences in the k-max to explain variations in an author's impact. Also, we investigate possible mechanisms that affect the k-max, both in the magnitude of the k and some possible strategies of increasing it, such as self-citations.

A personal network (ego network) intrinsically embeds social mechanisms and can generate different benefits to the focal node depending on its structure. Vacca (2020) explains that tightly-knit ego networks generate bonding social capital and may result in higher levels of cooperation and support among members. Unlike other networks (e.g. biological or social), in citation networks, the links cannot be severed; they are permanent, and the cost of maintaining the link (a citation) is zero once it appears. Given this particularity, we presume that a more cohesive inner core of a personal citation network could only benefit the focal author. Furthermore, the possible adverse effects of a highly cohesive network, such as greater social pressure or limits on individual freedom (Vacca 2020), are not a matter of concern.

Since the k-max of our citation networks contains papers that receive at least k citations, and these may come from different sources (self, collaborator, or third-party), we hypothesize that an author may partially have the ability to increase the k through self-citations or if collaborators reciprocate citations as well. Nevertheless, it is worth mentioning that we do not imply that all self or collaborator-citations are artificially boosting the impact or the k. For instance, an author could naturally have a large k if several of their papers are impactful within a small community and are usually co-cited.

We would expect differences between the citation networks of early and senior-career researchers and across fields, but it is not evident how they differ between genders, particularly the innermost core (k-max). The present paper aims at contributing to the research in how gender-differentiated patterns in their citation network can translate into permanence and promotion in academia.

Moreover, we use a Blinder-Oaxaca (BO) decomposition to explain further how the k-max and other variables contribute either positively or negatively to the citation gender gap, in the fashion of the wage gender gap literature. Thus, we examine how much of the gap can be explained by differences in observable characteristics or *endowments* (including the k-max) and how much is due to those characteristics having different effects on citations (*coefficients*). To our knowledge, this paper is the first, in the growing literature on gender bias in academia, to apply such a decomposition approach. Thus, this study not only contributes specifically to the gender bias literature in academia but may also inform policy-makers to design policies targeting gender equality.

The paper is organized as follows: section "The Mexican National System of Researchers" explains the Mexican National System of Researchers briefly; section "Data" presents the data and how the citation network is constructed; section "Network topology and impact" shows the results of the relationship between an author's networks topology and the number of citations; section "Network topology and impact" shows some measures that relate to the k-max and whether there are gender differences; section "Illustrative cases} illustrates particular researchers to better understand the dynamics of citation patterns and k-max. Finally, section "Conclusions" provides concluding comments.

The Mexican National System of Researchers

The National System of Researchers (*Sistema Nacional de Investigadores*, SNI) was created in 1984 and conceived initially to mitigate the acute income loss of faculty doing full-time research due to the economic crisis and aimed to support research activities across the country (Sandoval-Romero and Larivière 2020; Francisco et al. 2020). The evaluation process to enter the SNI and be promoted relies on peer review committees and assesses a mix of the number of publications and their impact. The SNI is divided into five levels in which members are classified: Candidates, SNI I, SNI II, SNI III, and

Table 1 SNI areas

Area	Name
1	Physics, Mathematics and Earth Sciences
2	Biology and Chemistry
3	Medicine and Health
4	Humanities and Behavioral Sciences
5	Social and Economic Sciences
6	Biotechnology and Agricultural Sciences
7	Engineering and Industry

Emeritus. There are seven evaluation committees depending on the researchers' area of expertise (Table 1).

The SNI provides a monthly monetary compensation determined by the Federal Government and depends solely on the level: Candidate receiving the lowest stimulus and Emeritus the highest. This compensation serves as a salary complement and represents, on average, 30% of the income but can represent up to 50% of it (Sandoval-Romero and Larivière 2020). This reward system has proven to incentivize the production of academic work. For instance, Rodríguez Miramontes et al. (2017) find that between 1991 and 2011, 83% of articles published by Mexican researchers were written by at least one member of the SNI in that period (Sandoval-Romero and Larivière 2020).

The evaluation periods are shorter for early-career researchers (Candidates and SNI I) and longer for seniors (SNIs II and III); thus, the highest rescission rates are within Candidate (41%) and SNI I (19.3%) (Sandoval-Romero and Larivière 2020). The first stages of the SNI remain the bottleneck overall, but especially for women since they are mainly represented in those levels (Appendix A).

Female presence in the SNI has increased but remains insufficient; in 2018, only 37% active members were women[1] (CONACYT 2018). However, the percentage of female researchers is heterogeneous across areas (Appendix A), ranging from 22% in Area 1 (Physics, Mathematics, and Earth Sciences) to 52% in Area 3 (Medicine and Health). Furthermore, comparing the percentage of female researchers across levels, we observe that it substantially decreases as the level increases, going from 44% in the lowest level (Candidate) to 23% in the highest one (SNI III)[2].

Data

We have access to public information on the researchers who were part of the Mexican National System of Researchers in 2018. This database contains 28,639 researchers, and we matched 11,039 authors to their corresponding information in the Microsoft Academic Graph (MAG) dataset[3] using their full names and institution. Even though the researchers in SNI are not randomly distributed and do not represent the whole population of researchers in Mexico, we verified that our matched sample has the same

[1] There were 28,639 active SNI members in 2018, but we were able to identify the exact gender of 27,775.

[2] Emeritus researchers (highest level) are not reported in the public SNI dataset.

[3] Information retrieved in August of 2021.

Table 2 Descriptive statistics

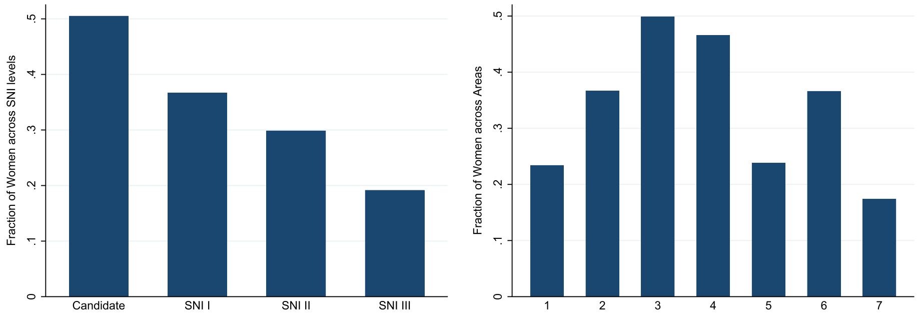
Variable	Statistic	All	Female	Male
Citations	Mean	602.2	522.7	642.3
	Std	877.8	623.0	979.3
Third-party citations	Mean	463.1	420.1	484.8
	Std	662.7	540.4	715.7
Collaborator citations	Mean	86.6	67.5	96.2
	Std	187.6	86.7	221.1
Self-citations	Mean	52.6	35.1	61.4
	Std	97.0	51.3	112.2
No. of publications	Mean	42.5	33.4	47.1
	Std	36.9	23.0	41.5
k-max	Mean	3.6	3.4	3.7
	Std	1.8	1.6	1.9
Rank of affiliation institution	Mean	8,964.7	8,763.7	9,066.1
	Std	1,628.0	1,567.4	1,649.0
Career length	Mean	19.0	17.8	19.5
	Std	7.9	6.9	8.2
Avg. co-authors per paper	Mean	6.3	6.2	6.3
	Std	30.0	10.1	36.0

characteristics as the SNI population, such as the proportion of women per area or proportion of women per SNI level (Appendix A). For each author, we retrieved from MAG the number of citations and publications and their institutional rank. An institution's rank is a measure constructed by MAG and roughly defined as the logarithm of the probability of an entity being "important", where importance is calculated using its relationships with other entities in the graph.

Considering that we are interested in the role of *k*-max as a determinant of success, we only kept those researchers with at least 100 citations in the (MAG) dataset; in this way, we obtained more dense citation networks. Researchers with fewer citations tend to have *k*-max with lower *k* because citation distributions tend to be highly skewed, where most authors have none or very few citations. As a result, our final baseline sample consists of 2363 researchers in all areas.

It is worth noting that our final sample is not representative of the whole population of SNI researchers since we are only considering those who have many citations (>100); however, since the authors we identify in MAG are representative of the SNI population, our sample of 2.3k authors should also be representative of the highly cited authors in the SNI population. Due to this, the distribution of our sample differs from the distribution of the population of SNI researchers across areas because different areas have different citation patterns. For instance, Ioannidis et al. (2019) find that the median of citations in General Arts, Humanities & Social Sciences is 28 while in Biology is 140 and Chemistry 129. We show in Appendix C frequency tables by area and SNI level of our sample.

As shown in Table 2, the mean number of citations in the sample is 602.2 citations, where women have on average 119.7 fewer citations than men, and this difference is statistically significant. We observe the same pattern when using third-party citations (women have 64.7 fewer citations than men) and collaborator citations. Women cite



(a) Levels

(b) Areas

Fig. 1 Share of women across SNI Levels and Areas

their papers less relative to men, consistent with the literature (King et al. 2017). The average number of publications is 42.5 papers, 33.4 for female researchers and 47.1 for male researchers. Moreover, women have on average a less interconnected k-max ($k = 3.4$) than men ($k = 3.7$).

Figure 1a presents the fraction of women in each of the four SNI levels: Candidate and levels I, II and III. There are 49% of men at the candidate level, as opposed to 51% of women. Thus, the higher the SNI level, the lower the representation of women. Medicine and Health (Area 3) is the area of knowledge with the largest fraction of women, followed by Humanities and Behavioral Sciences (Area 4), while Engineering and Industry (Area 7) has the lowest fraction (Fig. 1b).

Citation networks

For each author with at least 100 citations uploaded in MAG, we construct a citation network as follows: (i) we consider all articles where the author appears (ii) we consider all the articles that cite at least one article of the author in question. For simplicity, we remove the directionality of the link as we are only interested in the level of the network's interconnectivity. In this manner, the citation network is an undirected network where nodes represent articles and links represent citations from one article to another. We note that for every link, there must be at least one incident node representing an article of the focal author (we do not consider citations between two articles in which the focal author has no authorship) [4].

Next, we partition the nodes into three different classes: self, collaborator, and third-party author; depending on whether the author is part of the list of co-authors, a collaborator of the author is part of the list of co-authors, or none of the above, respectively. Classes are mutually exclusive, meaning that a node can only belong to one of them; if the paper represented by the node has the author and a collaborator as co-authors, we consider it a self-citation.

Finally, we obtain the largest k for which there exists a k-core using the graph-tool library (Peixoto 2014). This allows us to compute the proportion of nodes of each class in both the complete network and the main core.

[4] Only papers where we can identify the complete list of authors are considered; therefore, the citation network of an author could be made from a subset of their citations.

Network topology and impact

As explained above, an author's citation network structure can be seen as the result of citations coming from third-party authors, collaborators, and the researcher's self-citation behavior. Different citation behaviors could translate into a different network topology that may have direct and indirect effects on academic impact. Therefore, our variables of interest for each researcher are the number of citations, third-party citations, collaborator citations, and self-citations for all the papers published until August of 2021.

Our measure of network structure is the maximal value of *k* for which a *k*-core exists, calculated using the citation network of each author. Intuitively, the *k*-max of the citation network captures the innermost core of highly intercited papers. In our case, due to the construction of the network, the *k*-max should always contain papers authored by the SNI member we are considering. It is not sufficient to have the highest number of total citations to have a dense *k*-max (high *k*); several of the authors' papers must be co-cited simultaneously.

Our dependent variables are over-dispersed count variables that can be analyzed either by log-transforming them and then using Ordinary Least Squares or by using a Negative Binomial regression model without such transformation. We use both models and include controls such as productivity (logarithm of the number of papers), the logarithm of the rank of the affiliation institution, the researcher's area, level in the SNI, career length[5] and the logarithm of the average number of co-authors per paper.

Table 3 shows the results of our estimation. As shown, independently of the model employed, there is a positive and significant correlation between a higher *k* of the *k*-max and the researcher's citations (total, third-party, collaborator, and self-citations). In other words, if the citation network of an author has a highly interconnected subnetwork, it follows that the author is more likely to have more citations. Interestingly, the coefficient is smaller for third-party and collaborator citations but grows for self-citations. There is also a significant and positive correlation between the number of publications and any type of researcher's citations, while the correlation is negative and significant for the rank of the affiliation institution. This last result is consistent with previous findings that show a positive correlation between institutional prestige and the probability of becoming a top-cited scientist in the long run (Li et al. 2019).

If we consider the area of knowledge, it is observed that researchers in Medicine & Health (Area 3) and Humanities & Behavioral Sciences (Area 4) have fewer citations consistently when compared to Engineering & Industry (Area 7, the omitted category). This result contrasts with the results in Gonzalez-Brambila and Veloso (2007) who show that researchers in Health Sciences receive the largest number of citations per four years of all SNI areas among researchers that were part of the SNI from 1991 to 2002.

By SNI level, we see that researchers in the most prestigious level (SNI III) receive between 59% and 72% more citations relative to the mean of Candidates, and SNI II researchers between 48% and 52% more citations than Candidates. Finally, there is a positive and significant correlation between the average number of co-authors per paper

[5] Career length is calculated as the time between an author's first publication and 2021, the year at which the MAG was consulted.

Table 3 Effect of individual and network characteristics on ln(citations) of an author

Variables	**OLS**				**Negative Binomial**			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
	ln(citations)	**ln(third-party cit.)**	**ln (collaborator cit.)**	**ln (self-citations)**	**Citations**	**Third-party cit.**	**Collaborator cit.**	**Self-citations**
ln(*k*-max)	0.772***	0.605***	1.115***	1.759***	0.749***	0.616***	1.182***	1.743***
	(0.035)	(0.037)	(0.049)	(0.036)	(0.047)	(0.050)	(0.058)	(0.047)
ln(No. Publications)	0.350***	0.310***	0.384***	0.836***	0.300***	0.271***	0.329***	0.768***
	(0.027)	(0.029)	(0.036)	(0.028)	(0.049)	(0.052)	(0.044)	(0.048)
ln(Rank of affiliation institution)	−0.289***	−0.298***	−0.388***	−0.228***	−0.408***	−0.398***	−0.517***	−0.161***
	(0.073)	(0.081)	(0.106)	(0.068)	(0.085)	(0.095)	(0.104)	(0.060)
Area 1: Physics, Mathematics and Earth Sciences	−0.217***	−0.363***	0.331***	−0.064	−0.178**	−0.272***	0.240***	−0.110***
	(0.045)	(0.052)	(0.063)	(0.040)	(0.069)	(0.079)	(0.062)	(0.041)
Area 2: Biology and Chemistry	−0.054	−0.114**	0.248***	−0.007	−0.065	−0.103*	0.202***	−0.008
	(0.041)	(0.045)	(0.062)	(0.039)	(0.049)	(0.054)	(0.063)	(0.036)
Area 3: Medicine and Health	−0.117***	−0.113**	−0.242***	−0.227***	−0.159***	−0.149***	−0.238***	−0.169***
	(0.044)	(0.048)	(0.073)	(0.048)	(0.053)	(0.057)	(0.069)	(0.046)
Area 4: Humanities and Behavioral Sciences	−0.267***	−0.281***	−0.209*	−0.465***	−0.266***	−0.281***	−0.129	−0.323***
	(0.078)	(0.084)	(0.121)	(0.103)	(0.095)	(0.099)	(0.175)	(0.086)
Area 5: Social and Economic Sciences	−0.099	−0.089	−0.154	−0.457***	−0.029	−0.009	−0.124	−0.441***
	(0.087)	(0.092)	(0.122)	(0.096)	(0.092)	(0.096)	(0.116)	(0.076)
Area 6: Biotechnology and Agricultural Sciences	−0.097**	−0.113**	0.029	−0.058	−0.116**	−0.118**	−0.015	−0.062*
	(0.041)	(0.046)	(0.061)	(0.039)	(0.046)	(0.051)	(0.059)	(0.035)
Level I	−0.007	0.017	0.105	−0.090	−0.008	−0.011	0.121	−0.046
	(0.068)	(0.078)	(0.094)	(0.060)	(0.089)	(0.100)	(0.094)	(0.057)
Level II	0.257***	0.318***	0.309***	−0.034	0.326***	0.345**	0.320***	0.043
	(0.077)	(0.088)	(0.106)	(0.070)	(0.126)	(0.138)	(0.107)	(0.071)
Level III	0.470***	0.556***	0.487***	−0.072	0.627***	0.668***	0.587***	0.084
	(0.091)	(0.103)	(0.123)	(0.084)	(0.146)	(0.157)	(0.143)	(0.103)
Career length	0.008***	0.010***	0.004	−0.011***	0.004*	0.005**	0.005	−0.009***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
ln(Avg. co-authors per paper)	0.340***	0.361***	0.589***	−0.063**	0.342***	0.355***	0.584***	−0.110***
	(0.036)	(0.037)	(0.064)	(0.029)	(0.043)	(0.046)	(0.063)	(0.032)
Constant	5.765***	5.842***	3.330***	0.655	7.249***	7.160***	4.918***	0.409
	(0.676)	(0.742)	(0.968)	(0.623)	(0.798)	(0.895)	(0.975)	(0.559)
lnalpha					−1.120***	−0.898***	−0.551***	−1.582***
					(0.054)	(0.048)	(0.037)	(0.062)
Observations	2363	2362	2334	2311	2363	2363	2363	2363

Robust standard errors in parentheses

Ommited dummies are Candidate and Engineering and Industry

*** $p < 0.01$

** $p < 0.05$

$*p < 0.1$

and total, third-party and collaborator citations, but the relationship becomes negative for self-citations, independently of the model used.

We show in section "Data" that there are significant differences or gaps across female and male researchers. Considering the results above, we can answer the differential effects of the different determinants on the academic impact of female and male researchers. We use the logarithm of the total number of a researcher's citations as a measure of academic impact. Following the seminal works of Oaxaca and Blinder (1973, 1973), we use the Blinder-Oaxaca (BO) decomposition method to study the citation gap. This method has been widely applied in economics to study gender/racial wage gaps.

We first estimate a two group-specific regression model (see Table 13 in Appendix D) and then perform the decomposition. As shown in Table 4, the decomposition output reports the mean predictions of the logarithm of citations for men and women and their difference. In our sample, the mean of ln(citations) is 5.976 for men and 5.882 for women, yielding a citation gap of 0.0940. The citation gap is divided into three parts in the first column of the decomposition output (*endowments*, *coefficients* and *interaction*).

The first term reflects the mean increase in women's citations if they had the same characteristics as men (effects due to women having different *endowments*). The increase of 0.142 indicates that differences in productivity (logarithm of the number of papers), the logarithm of the rank of the affiliation institution, the researcher's area, level in the SNI, career length, and the logarithm of the average number of co-authors per paper account for about 151% the citation gap.

The second term quantifies the change in women's citations when applying the men's coefficients to the women's characteristics (effects due to those characteristics having different influences on citations-*coefficients*). The overall difference in citations decreases when applying men's coefficients. As shown in Table 12 of Appendix D, the average number of co-authors per paper, career length, and a higher level of SNI (relative to Candidate) boost cites more for women than for men (i.e., the coefficients are greater for women), which explains why when applying the coefficients of men to these factors (column 3 of Table 4) the gap decreases. The third term is the interaction term that measures the simultaneous effect of differences in endowments and coefficients, and is not significant. Overall, these results show that differences in endowments between women and men explain the citation gap.

Gender differences in *k*-max

In section "Network topology and impact", we explore the relationship between various determinants of an author's characteristics and the academic impact. In particular, we find that a higher *k* of the *k*-max author's citation network does have a positive and significant association with the number of citations. Thus, if a less interconnected core produces fewer citations, we would like to know the extent of these gender differences.

In Fig. 2, we present the probability density function of the maximal *k* values for which there is a *k*-core (*k*-max), distinguishing between women and men[6]. We observe that

[6] We approximate the density f(*k*) from observations on *k* using an Epanechnikov kernel density estimator with width one due to the integer nature of *k*. It is a continuous approximation of the empirical distribution of *k*, and in our case, this density is almost identical if we employ a Gaussian kernel.

Table 4 Blinder–Oaxaca decomposition

Variables	(1) Overall	(2) Endowments	(3) Coefficients	(4) Interaction
ln(*k*-max)		0.0446***	0.192**	0.0117*
		(0.0136)	(0.0837)	(0.00633)
ln(No. Publications)		0.0934***	0.0896	0.00762
		(0.0151)	(0.178)	(0.0152)
ln(Rank of affiliation institution)		−0.0123**	1.456	0.00529
		(0.00586)	(1.360)	(0.00554)
Area 1: Physics, Mathematics and Earth Sciences		−0.00929	−0.0110	−0.00716
		(0.00724)	(0.0130)	(0.00847)
Area 2: Biology and Chemistry		0.000778	−0.0131	0.00170
		(0.00349)	(0.0288)	(0.00433)
Area 3: Medicine and Health		0.0142	−0.00586	0.00289
		(0.0103)	(0.0249)	(0.0122)
Area 4: Humanities and Behavioral Sciences		0.00641**	0.00589	−0.00248
		(0.00320)	(0.00644)	(0.00299)
Area 5: Social and Economic Sciences		−0.00436	0.00619	0.00380
		(0.00418)	(0.00536)	(0.00406)
Area 6: Biotechnology and Agricultural Sciences		0.000222	−0.0248	0.00315
		(0.00270)	(0.0206)	(0.00394)
Level I		−0.00103	−0.0215	0.00277
		(0.00778)	(0.0944)	(0.0124)
Level II		0.0127	−0.0121	−0.00223
		(0.00797)	(0.0382)	(0.00821)
Level III		0.0368***	−0.00344	−0.00388
		(0.0132)	(0.0135)	(0.0138)
Career length		0.0216**	−0.107	−0.0107
		(0.00860)	(0.0778)	(0.00876)
ln(Avg. co-authors per paper)		−0.0617***	−0.0951	0.00949
		(0.0116)	(0.100)	(0.0105)
Men	5.976***			
	(0.0243)			
Women	5.882***			
	(0.0303)			
Difference	0.0940**			
	(0.0398)			
Endowments	0.142***			
	(0.0296)			
Coefficients	−0.0701**			
	(0.0276)			
Interaction	0.0220			
	(0.0197)			
Constant			−1.526	
			(1.421)	
Observations	2363	2363	2363	2363

Bootstrapped standard errors in parentheses (100 replications)

Ommited dummies are Candidate and Engineering and Industry

***$p < 0.01$

**$p < 0.05$

*$p < 0.1$

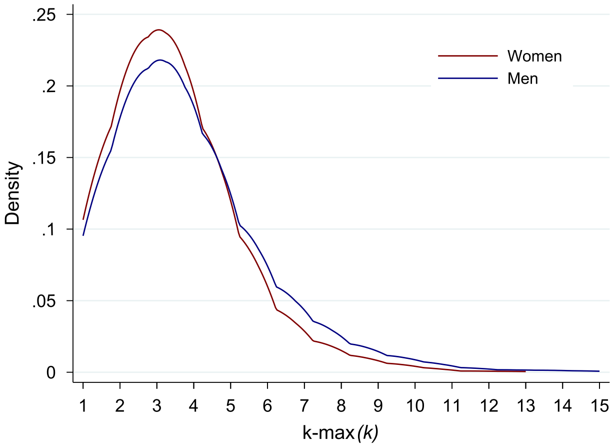


Fig. 2 Probability density distribution of *k*-max by gender

women tend to have less interconnected cores in their citation networks, the average *k* for women is 3.4 while for men is 3.7, and the maximum value of *k* for women (13) is smaller than for men (15).

Conducting a Kolmogorov-Smirnov test (Appendix B), we find evidence that the *k*-max probability density distribution functions of women and men are not equal and that women tend to have a less interlinked inner core of their citation networks than men. Thus, if higher values of *k* are associated with more citations and the advantages this entails for an author, it would mean that women would not be benefiting as much as men through this channel, confirming the results of the Blinder-Oaxaca decomposition (Table 4).

If women have less interlinked citation networks, we would like to investigate possible determinants of a higher *k* in the *k*-max and whether women and men display different patterns in these. Table 5 show the correlation matrices of the *k*-max magnitude and the number of nodes and edges in both the whole network and the *k*-max subnetwork for women and men, respectively. For both genders, we find a positive relationship between the number of edges in the network and the *k*-max, stronger for men (0.591) than for women (0.431). Also, there exists a positive relationship between the number of nodes in the network and the *k*-max, stronger for men than for women (0.503 and 0.302, correspondingly). However, we find that more interlinked cores tend to have fewer nodes and more edges in the citation network.

Considering the type of citations an author can receive (self, collaborator and third-party citations), we analyze how these correlate with the *k*-max. Our interest is to shed light on whether an author's strategic behavior through the increment in self-citations and collaborator-citations can increase the *k* magnitude. In Table 6, we show the correlation matrices of the *k*-max and the share of each type of citation out of the total citations. We find that self-citations have the strongest positive correlation coefficient with *k*-max and collaborator-citations to a lesser extent. Interestingly, we find that third-party citations are negatively correlated with *k*-max, which may indicate that authors able to gather more self and collaborator-citations tend to have more total citations ultimately.

Table 5 Correlation matrix of nodes, edges and *k*-max

	# of nodes: network	# of edges: network	# of nodes: core	# of edges: core	*k*-max
WOMEN					
# of nodes: network	1				
# of edges: network	0.978***	1			
# of nodes: core	−0.0368	−0.0390	1		
# of edges: core	0.134***	0.213***	0.802***	1	
k-max	0.304***	0.436***	−0.264***	0.172***	1
MEN					
# of nodes: network	1				
# of edges: network	0.970***	1			
# of nodes: core	0.0640*	0.0504*	1		
# of edges: core	0.378***	0.427***	0.816***	1	
k-max	0.503***	0.589***	−0.137***	0.291***	1

*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

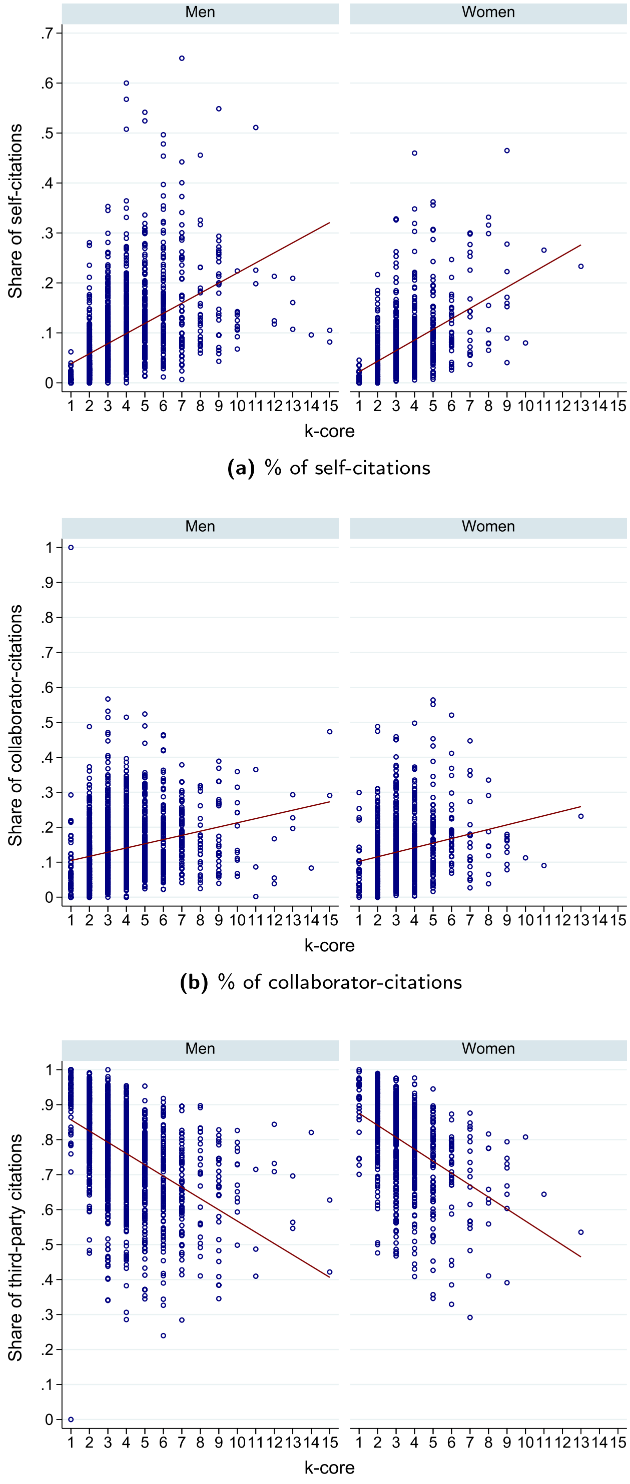
Table 6 Correlation matrix of type of citations and *k*-max

	% Self-citations	% Collaborator-citations	% Third-party citations	*k*-max
WOMEN				
% Self-citations	1			
% Collaborator-citations	0.142***	1		
% Third-party citations	−0.655***	−0.841***	1	
k-max	0.496***	0.219***	−0.438***	1
MEN				
% Self-citations	1			
% Collaborators-citations	0.137***	1		
% Third-party citations	−0.716***	−0.789***	1	
k-max	0.473***	0.247***	−0.467***	1

% of each type with respect to total citations
*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

In Fig. 3 we show *k*-max and type of citations in our data, considering the share of each type with respect to the total number of citations. As seen in Table 6, there is a positive correlation between the largest *k* for which there is a *k*-core and self-citations and between the largest *k* and collaborator-citations, with a higher slope for the first one than for the second one. On the contrary, there is a negative association between the largest *k* and third-party citations. However, there is no visible difference of the correlation coefficients between women and men.

We explore further whether, given a *k*-max, there are differences between genders on the type of cite (share) they receive. Figure 4 shows the median of each share of citations differentiated by gender, for each *k*. Strikingly, for less dense citation networks ($k < 3$), there are no observable differences between women and men for any type of cite.

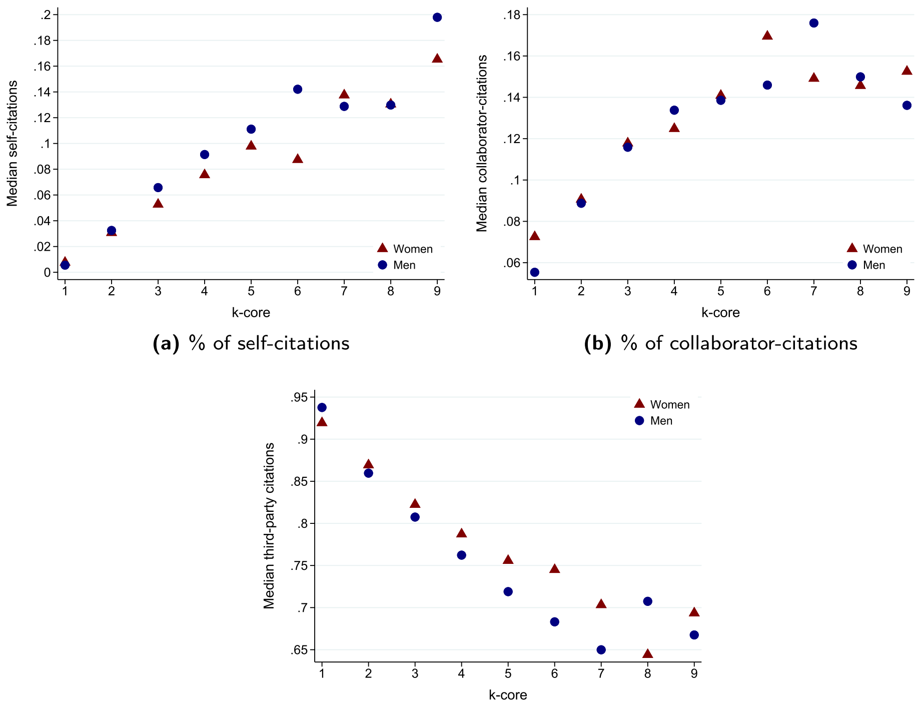


(a) % of self-citations

(b) % of collaborator-citations

(c) % of third-party citations

Fig. 3 Relation between *k*-max and type of citations



(a) % of self-citations

(b) % of collaborator-citations

(c) % of third-party citations

Fig. 4 Type of citations according to *k*-max and gender: median. *Notes*: we plot until $k = 9$ since it is the last *k* that contains at least two observations per gender

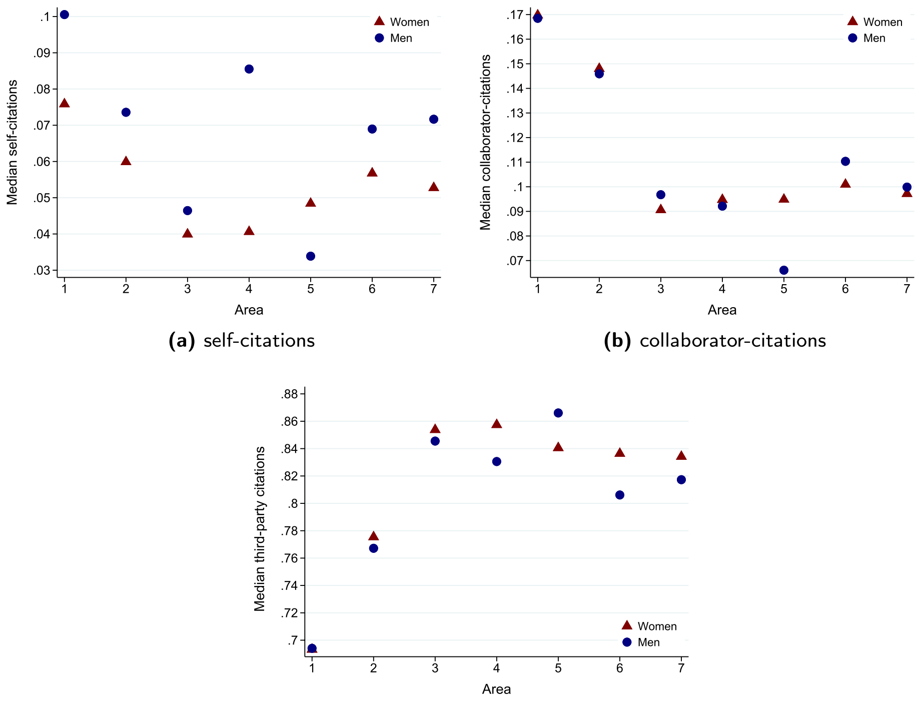
However, as *k* increases, the gap between genders widens. Women tend to have more third-party citations for each *k*. Men lean towards higher shares of self and collaborator-citations, for each *k*, suggesting a higher probability of reciprocating citations.

One may argue that some areas tend to have more collaborators and many papers, leading to more interconnected citation networks. Thus, we explore if, given an area, there are observable gender differences in the median of each share of citations differentiated by gender (Fig. 5). Area 1 (Physics, Mathematics, and Earth Sciences) has the highest share of self and collaborator-citations and the lowest third-party citations, more for men than women. On the contrary, Area 5 (Social and Economic Sciences) has the lowest share of self and collaborator-citations and the highest share of third-party citations. Therefore, there are marked differences across fields in the share of type of citations that could translate in disparities in success between women and men, both within Areas and between them.

We do not argue that all self and collaborator-citations artificially boost the magnitude of the *k*-max. For instance, an author with many papers or several collaborators benefits simply due to that. Still, this is a finding that would be worth further exploring.

Illustrative cases

This section presents an illustrative example of the relationship between highly interconnected citation networks and the number of citations by collaborators and third-party authors. We propose comparing two researchers with different citation networks, looking into their citation patterns discussed in the previous sections.



(a) self-citations

(b) collaborator-citations

(c) third-party citations

Fig. 5 Type of citations according to area and gender: median

Researchers A and B are both physicists working in materials science and have 6,625 and 4,178 citations across 112 and 227 articles, respectively. The largest k for which there is a k-core for Researcher A is 8, while this value is 5 for Researcher B. Thus, A has more citations and a more interconnected citation network than B who has fewer citations and a lower k-core. In Fig. 6A, we can see the difference in the interconnectivity citation network of both researchers: Researcher A has a more densely connected core than Researcher B.

The k-core decomposition of both authors shows that researcher A presents a more dense network for various levels of k, meaning that articles that cite their work usually cite several of their articles. Interestingly, the distribution of the number of citations for researcher A, while higher, is more compact, with fewer outliers than Researcher B (Fig. 6 B. Researcher B has only 25% of their articles with more than 13 citations, while researcher A has 42%. However, the standard deviation of the number of citations of researchers A and B is 62 and 76, respectively, with researcher B having more outlier articles. Moreover, researcher A publishes more often in less ranked venues than researcher B, even though they share the same field (Fig. 6C). On average, researcher B publishes in journals with rank 7,685 while researcher A in journals with rank 10,268.

Finally, the source of citations of both researchers is quite different. While 82.6% of citations come from third-party authors for B, A only receives 59% (Fig. 6A). Researcher A receives 31.8% of their citations from collaborators while researcher B receives only 13.8%, less than half the proportion by researcher A. The same pattern occurs for self-citations: Researcher A has 8.9% while Researcher B 3.5%. These findings are consistent with our

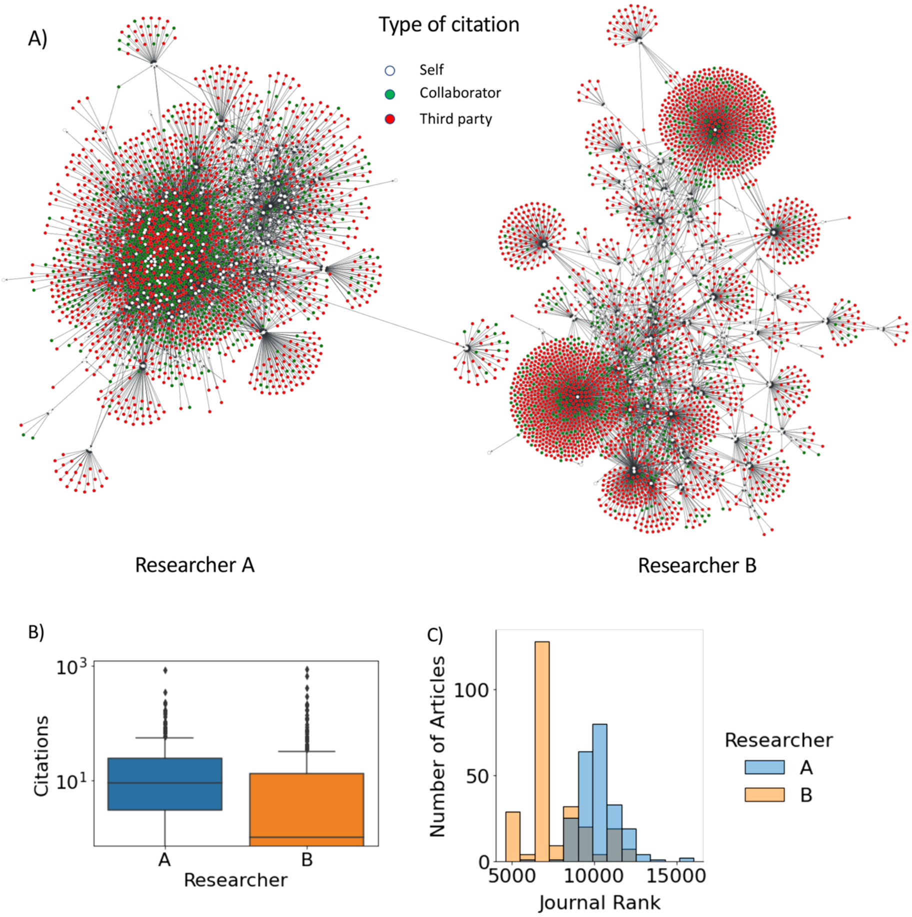


Fig. 6 Example of networks with different *k*-max. **A** Citation networks differs structurally. Researcher A has a 8-core while researcher B has a 5-core. **B** Distribution of citations across papers of both researchers. **C** Distribution of the ranks (from MAG) of the journals where researchers A and B publish

findings in the previous sections where a more interconnected citation network core positively correlates with self and collaborator-citations and negatively with third-party citations.

Conclusions

In this work, we study the effect of the size of interconnected nodes of an author's citation network and the number of citations the authors receives. We find a positive relationship between the size of the main core in the citation network of an author (magnitude of *k* of the *k*-max), a proxy for the size of their interlinked articles, and their number of citations.

We observe that more interlinked citation networks correlate with a large share of self and collaborator-based citations, and a low share of third-party citations as a percentage of the total number of citations. We argue that this could serve as a mechanism to directly or indirectly boost citations, with the caveat that it could also occur naturally in some research areas and exceptional cases. For instance, there are notable differences across fields in the share of citation types in which Area 1 (Physics, Mathematics, and Earth Sciences) has the highest share of self and collaborator-based citations, as well the lowest third-party citations, more for men than women.

We show a statistically significant difference between the level of the interconnectivity of men and women citation networks, where the women tend to have a less interlinked inner cores. Thus, if women tend to have consistently less interlinked citation networks, this could limit the permanence and promotion of women careers.

We also explore the citation gender gap through a Blinder-Oaxaca (BO) decomposition. We examine how much of the gap can be explained by differences in observable characteristics or *endowments* (including *k*-max) and how much is due to those characteristics having different effects on citations (*coefficients*). Our results show that differences in endowments between women and men explain much of the citation gap.

In this sense, further research could explore how citation reciprocity differs between genders and how the topology of the citation network evolves with the career. These would shed light on how strategic behavior, other than self-citations, affects an academic career and whether we can find significant gender-differentiated determinants. We do not affirm that all self and collaborator-citations artificially boost the *k*-max. For instance, an author with many papers or several collaborators benefits simply due to that. However, this is a finding that would be worth further exploring.

Appendix A: SNI population and matched sample

The matched sample corresponds to all SNI researchers we identified in the MAG data through the normalized name and institution (Tables 7, 8, 9, 10, 11, 12, 13).

Table 7 Number of researchers in each Area: matched sample and population

Area	Matched sample		Population	
	Number	Share	Number	Share
1	1489	0.14	4199	0.15
2	1367	0.12	3469	0.12
3	2192	0.20	4602	0.17
4	1364	0.12	4371	0.16
5	1874	0.17	4262	0.15
6	1697	0.15	4137	0.15
7	1021	0.09	2735	0.10
Total	11,004	1.00	27,775	1.00

Table 8 Number of women and men in each area: matched sample and population

Area	Matched sample				Population			
	Women	Share	Men	Share	Women	Share	Men	Share
1	611	0.41	878	0.59	1769	0.42	2430	0.58
2	467	0.34	900	0.66	1266	0.36	2203	0.64
3	844	0.39	1348	0.61	1841	0.40	2761	0.60
4	303	0.22	1061	0.78	966	0.22	3405	0.78
5	909	0.49	965	0.51	2113	0.50	2149	0.50
6	354	0.21	1343	0.79	935	0.23	3202	0.77
7	505	0.49	516	0.51	1424	0.52	1311	0.48
Total	3993	0.36	7011	0.64	10,314	0.37	17,461	0.63

Table 9 Number of women and men in each SNI level: matched sample and population

Level	Matched sample				Population			
	Women	Share	Men	%	Women	Share	Men	Share
Candidate	930	0.45	1138	0.55	2832	0.44	3624	0.56
SNI I	2327	0.37	3932	0.63	5524	0.38	9090	0.62
SNI II	586	0.31	1280	0.69	1441	0.33	2984	0.67
SNI III	150	0.18	661	0.82	517	0.23	1763	0.77
Total	3993	0.36	7011	0.64	10,314	0.37	17,461	0.63

Appendix B: Kolmogorov–Smirnov test for equality of *k*-max distribution functions

Table 10 Kolmogorov–Smirnov test for equality of *k*-max distribution functions

Smaller group	Largest difference	*p*-value
Men	0.0066	0.955
Women	−0.0734	0.003
Combined K-S	0.0734	0.007

Table 11 Number of women and men in each Area: sample

Area	Women	Share	Men	Share	Total
1	84	0.23	275	0.77	359
2	230	0.37	397	0.63	627
3	202	0.50	203	0.50	405
4	34	0.47	39	0.53	73
5	20	0.24	64	0.76	84
6	153	0.37	265	0.63	418
7	69	0.17	328	0.83	397
Total	792	0.34	1571	0.66	2363

Appendix C: Sample by SNI Area and Level

Table 12 Number of women and men in each Level: sample

Level	Women	Share	Men	Share	Total
Candidate	50	0.51	49	0.49	99
SNI I	506	0.37	874	0.63	1380
SNI II	186	0.30	437	0.70	623
SNI III	50	0.19	211	0.81	261
Total	792	0.34	1571	0.66	2363

Appendix D: Oaxaca decomposition

Table 13 Effect of individual and network characteristics on ln(citations) of an author by gender

Variables	**(1)** **ln(citations)**	**(2)** **ln(citations)**
	Men	Women
ln(k-max)	0.818***	0.648***
	(0.0415)	(0.0611)
ln(No. Publications)	0.359***	0.332***
	(0.0340)	(0.0443)
ln(Rank of affiliation institution)	−0.212**	−0.373***
	(0.0880)	(0.129)
Area 1: Physics, Mathematics and Earth Sciences	−0.238***	−0.135
	(0.0495)	(0.104)
Area 2: Biology and Chemistry	−0.0657	−0.0206
	(0.0482)	(0.0877)
Area 3: Medicine and Health	−0.136**	−0.113
	(0.0553)	(0.0876)
Area 4: Humanities and Behavioral Sciences	−0.217*	−0.354***
	(0.119)	(0.109)
Area 5: Social and Economic Sciences	−0.0361	−0.281
	(0.0990)	(0.185)
Area 6: Biotechnology and Agricultural Sciences	−0.138***	−0.00906
	(0.0483)	(0.0885)
Level I	−0.0212	0.0124
	(0.0987)	(0.0936)
Level II	0.242**	0.293***
	(0.109)	(0.109)
Level III	0.463***	0.517***
	(0.119)	(0.159)
Career length	0.00620***	0.0122***
	(0.00230)	(0.00361)
ln(Avg. co-authors per paper)	0.320***	0.379***
	(0.0459)	(0.0481)
Constant	5.033***	6.560***
	(0.824)	(1.192)
Observations	1,571	792
R-squared	0.606	0.536

Robust Standard errors in parentheses

Ommited dummies are Candidate and Engineering and Industry

***p<0.01

**p<0.05

*p<0.1

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 14 December 2021 Accepted: 4 August 2022
Published online: 19 August 2022

References

Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2005) Large scale networks fingerprinting and visualization using the k-core decomposition. Advances in Neural Information Processing Systems, pp 41–50

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8:436–455. https://doi.org/10.2307/144855

Burleson-Lesser K, Morone F, Tomassone MS, Makse HA (2020) K-core robustness in ecological and financial networks. Sci Rep 10(1):3357

CONACYT (2018) Informe general del estado de la ciencia, la tecnología y la innovación. Technical report

Campos-Vázquez RM, Lustig N, Scott J (2018) Inequality in Mexico 5/2018

Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J (2021) k-core genes underpin structural features of breast cancer. Sci Rep 11(1):1–17

Dorantes-Gilardi R, Ramírez-Álvarez AA, Terrazas-Santamaría D (2021) Is there a differentiated gender effect of collaboration with supercited authors? evidence from early-career economists. Technical report, El Colegio de México, Centro de Estudios Económicos

Fortunato S, Bergstrom CT, Bšrner K, Evans JA, Helbing D, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barab A-l (2018) Science of science. Science 359(6379). https://doi.org/10.1126/science.aao0185

Francisco J, Fontes G, Martí J, Antón MG, Author C, Francisco J, Fontes G (2020) The emergence of the new Mexican academic meritocracy. Higher Educ Govern Policy 1(2):138–151

Giatsidis C, Thilikos DM, Vazirgiannis M (2011) Evaluating cooperation in communities with the k-core structure. In: 2011 International conference on advances in social networks analysis and mining, pp 87–93. https://doi.org/10.1109/ASONAM.2011.65

Gonzalez-Brambila C, Veloso FM (2007) The determinants of research output and impact: a study of mexican researchers. Res Policy 36:1035–1051. https://doi.org/10.1016/j.respol.2007.03.005

Huang J, Gates AJ, Sinatra R, Barabási AL (2020) Historical comparison of gender inequality in scientific careers across countries and disciplines. Proc Natl Acad Sci USA 117(9):4609–4616. https://doi.org/10.1073/pnas.1914221117 (**1907.04103**)

Ioannidis JPA, Baas J, Klavans R, Boyack KW (2019) A standardized citation metrics author database annotated for scientific field. PLoS Biol 17(8):1–6

Kacem A, Flatt JW, Mayr P (2020) Tracking self-citations in academic publishing. Scientometrics 123(2):1157–1165. https://doi.org/10.1007/s11192-020-03413-9

King MM, Bergstrom CT, Correll SJ, Jacquet J, West JD (2017) Men set their own cites high: gender and self-citation across fields and over time. Socius 3:2378023117738903

Kong YX, Shi GY, Wu RJ, Zhang YC (2019) k-core: theories and applications. Phys Rep 832:1–32. https://doi.org/10.1016/j.physrep.2019.10.004

Li W, Aste T, Caccioli F, Livan G (2019) Reciprocity and impact in academic careers. EPJ Data Science 8:1–15. https://doi.org/10.1140/epjds/s13688-019-0199-3

Li W, Aste T, Caccioli F, Livan G (2019) Early coauthorship with top scientists predicts success in academic careers. Nat Commun 10:5170. https://doi.org/10.1038/s41467-019-13130-4

Lloyd M (2018) El sector de la investigación en México: entre privilegios, tensiones y jerarquías. Rev de la Educación Superior 47(185):1–31

Oaxaca R (1973) Male-female wage differentials in urban labor markets. Int Econ Rev 14:693–709. https://doi.org/10.2307/2525981

Paraskevopoulos P, Boldrini C, Passarella A, Conti M (2021) The academic wanderer: structure of collaboration network and relation with research performance. Appl Netw Sci 6(1). https://doi.org/10.1007/s41109-021-00369-4

Peixoto TP (2014) The graph-tool python library. figshare

Rodríguez Miramontes J, González Brambila CN, Maqueda Rodríguez G (2017) El sistema nacional de investigadores en México: 20 años de producción científica en las instituciones de educación superior (1991–2011). Investigacion Bibliotecologica 2017(Special Issue), 187–219. https://doi.org/10.22201/iibi.24488321xe.2017.nesp1.57890

Sandoval-Romero V, Larivière V (2020) The national system of researchers in Mexico: implications of publication incentives for researchers in social sciences. Scientometrics 122(1):99–126. https://doi.org/10.1007/s11192-019-03285-8

Seidman SB (1983) Network structure and minimum degree. Soc Netw 5(3):269–287. https://doi.org/10.1016/0378-8733(83)90028-X

Vacca R (2020) Structure in personal networks: constructing and comparing typologies. Netw Sci 8(2):142–167. https://doi.org/10.1017/nws.2019.29

Van Noorden R, Singh Chawla D (2019) Policing self-citations. Nature 572(7771):578–579

Van Der Wal JEM, Thorogood R, Horrocks NPC (2021) Collaboration enhances career progression in academic science, especially for female researchers. Proc R Soc B: Biol Sci 288(1958):1–10

Van den Besselaar P, Sandström U (2016) Gender differences in research performance and its impact on careers: a longitudinal case study. Scientometrics 106(1):143–162. https://doi.org/10.1007/s11192-015-1775-3

Wallace ML, Larivière V, Gingras Y (2012) A small world of citations? The influence of collaboration networks on citation practices. PLoS ONE 7(3):33339

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.