PhD students in life sciences can benefit from team cohesion

Zsófia Viktória Vida¹, István Péter Járay², Balázs Lengyel²-⁴

¹Department of Science Policy and Scientometrics, Library and Information Centre of the Hungarian Academy of Sciences, Budapest, 1051, Hungary
²Centre for Economic and Regional Studies, Eötvös Loránd Research Network, Budapest, 1097, Hungary
³Centre for Advanced Studies, Budapest Corvinus University, Budapest, 1093, Hungary
⁴Agglomeration and Social Networks Lendület Research Group, Budapest, 1097, Hungary

Abstract

Background: Scientific progress during doctoral studies is a combination of individual effort and teamwork. A recently growing body of interdisciplinary literature has investigated the determinants of early career success in academia, in which learning from supervisors and co-authors play a great role. Yet, it is less understood how collaboration patterns of the research team, in which the doctoral student participates, influences the future career of students. Here we take a social network analysis approach to investigate this and define the research team as the co-authorship network of the student.

Methods: We use the Hungarian Scientific Bibliography Database, which includes all publications of PhD students who defended theses from the year 1993. The data also include thesis information, and the publications of co-authors of students. Using this data, we quantify cohesion in the ego-network of PhD students, the impact measured by citations received, and productivity measured by number of publications. We run multivariate linear regressions to measure the relation of network cohesion, and publication outputs during doctoral years with future impact.

Results: We find that those students in life sciences, but not in other fields, who have a cohesive co-author network during studies and two years after defence receive significantly more citations in eight years. We find that the number of papers published during PhD years and closely after the defence correlates negatively while the impact of these papers correlates positively with future success of students in all fields.

Conclusions: These results highlight that research teams are effective learning environments for PhD students where collaborations create a tightly knit knowledge network.
Keywords
co-authorship network, academic success, scientific impact, scientific career

This article is included in the Research on Research, Policy & Culture gateway.
**Introduction**

As international competition in science accelerates, there has been a growing interest in the determinants of individual success in academia (Sinatra et al., 2016; Clauset et al., 2017; Fortunato et al., 2018). A general notion is that success breeds success because recognized publications open new opportunities for funding and collaboration (Bol et al., 2018). This has directed attention to young scholars because achievements at early stage might generate future success (Wang et al., 2019). The relation between scientific collaboration and success has been investigated for some time (de Solla Price & Beaver, 1966; Luukkonen et al., 1993; Melin & Persson, 1996; Katz & Martin, 1997; Sonnenwald, 2007; Hood & Wilson, 2001; Sarigöll et al., 2014). It is especially useful to understand early career success because young scholars need mentors to learn from and are more likely to stand out in case they work with successful supervisors (Sekara et al., 2018; Ma et al., 2020; Li et al., 2019). Although early careers might learn from more people, it is less understood how future success of early career researchers depend on the team they work with during their early career.

Teams in scientific research are gaining dominance across all fields (Wuchty et al., 2007; Ziman, 1994). Research teams typically include postdocs, graduate and undergraduate students who collaborate with the principal investigator and other seniors of the group (Mali et al., 2012). Such collaborations can be observed in co-authorship that is frequently used to map collaboration networks in teams (Beaver, 2001; Glänzel & Schubert, 2004) and are thought to influence success of projects (Uzzi & Spiro, 2005). Two counteracting mechanisms are important in this respect. On the one hand, projects can create more novelty in the case where co-authors have not collaborated before because they can combine diverse expertise (De Vaan et al., 2015; Vedres, 2017; Zeng et al., 2021). On the other hand, team cohesion generated by shared co-authors, strong and persistent collaboration, trust and previous success can provide an environment in which knowledge sharing are efficient (Uzzi & Spiro, 2005; Aral & Van Alstyne, 2011; Mukherjee et al., 2019). Thus, the question, whether early career researchers benefit more from diverse than from cohesive teams, is important because striving for novelty in scientific research and efficient learning are difficult to achieve at the same time.

In this paper, we take a social network analysis approach to investigate co-authorship networks of early career researchers. To quantify diversity and coherence in the collaboration network of students and across the author teams they belong to, we apply the network constraint measure developed by Burt (1992, 2000, 2001). This measure ranges between 0 and 1: high values around 1 show that co-authors of the PhD student work frequently together; low values around 0 show that the PhD student works with co-authors who are otherwise not collaborating with each other. This measure has been very widely used to capture diverse knowledge access through connections in various contexts including creative industries (Juhász et al., 2020), innovation (Tóth & Lengyel, 2021) and to capture the role of network cohesion in knowledge transfer (Reagens & McEvily, 2003; Tortoreillo et al., 2012).

Our empirical case concerns researchers who have had a successful defense in any Hungarian doctoral school between 1993 and 2010. Our data contains information on the dissertation, including the scientific field and year of defense, and bibliometric information data comes from publication records of egos and their co-authors. We estimate the accumulated number of citations at the eighth year following defense, which gives us a simple measure of success at the end of the early phase of academic career (Van Balen et al., 2012).

Cross-sectional linear regressions with year and scientific field dummies show that the number of papers published until the second year after the defense correlates negatively with accumulated citations, but the impact of these papers correlates strongly with future impact. This finding indicates that thorough work focusing on a few but important papers is the second year after the defense correlates negatively with accumulated citations, but the impact of these papers correlates strongly with future impact. These latter two co-efficients are not significant for other scientific fields. These results provide new evidence that PhD students in life-sciences can benefit from working in a cohesive research team probably because these provide a good learning environment.

**Methods**

**Ethical approval**

The ad-hoc ethical committee of the Library and Information Centre, Hungarian Academy of Sciences discussed the use of personal data for the purpose of the research project “Team cohesion of PhD students in life sciences”, and approval to use the data was given. Consent from the participants to use their data was waived by the committee. The Library and Information Centre of the Hungarian Academy of Sciences as data handler of the Hungarian Scientific Bibliography Database is authorized to use the data for scientific research as per law 1994/XL and 10/2021.(II. 24.) regulation decreed by the President of the Hungarian Academy of Sciences.

All other data used in the study was publicly available.
Data

We combine two data sources to collect information about early-career scholars. Data on doctoral defenses have been collected from www.doktori.hu, an openly available collection of all successful PhD theses defended in Hungarian doctoral schools starting from 1993, the year when the PhD system was introduced in the country. We downloaded data from the website in January 2017. This data contains 16,151 Hungarian PhD students who defended their theses until that date. Information provided includes the ID and name of every PhD student, the title of their thesis, the year of defense, scientific area, and the name of supervisors. Our second data source is the Hungarian Scientific Bibliography Database (MTMT) that contains the scientific publications’ metadata of all active Hungarian researchers.

The two databases can be matched on the individual student level. A total of 23% of students were matched between the doktori.hu database and MTMT data using their student IDs. The rest of the students were matched by hand using name and scientific field. We could identify 60% of the PhD students in the MTMT data. The number of PhD students who could be matched with an MTMT profile was 9,415 and is illustrated in Figure 1A by the year of defense. In our regression exercise, we focus on the future impact of PhD thus restrict the analysis to the 2,061 PhD students who defended theses in the 1993-2010 period.

**Figure 1. PhD students and publications in the data.** A. The number of PhD students by year in www.doktori.hu (light blue) and the successfully identified PhD students in the MTMT database (navy blue). Data includes all students who defended between 1993 and 2017 but the analysis will focus on those who defended in the 1993-2010 period. B. The number of publications by the Hungarian PhD holders (year of defenses between 1993-2010) and their co-authors between 1990 and 2019.
Bibliometric data was downloaded from MTMT by the data handler colleague of the Hungarian Academy of Sciences after the identification of PhD students in MTMT. This happened in two steps. First, we downloaded all 272,954 publication records of the identified 9,415 PhD students in 2017. Then, we identified 20,139 co-authors of PhD students in MTMT and downloaded their publication records in 2020. This final bibliometric dataset contains records of 1,205,184 papers published by 43,485 authors between 1990-2019. Note that only those authors are included who are affiliated with Hungarian institutions and must have registered on MTMT. There are around 50,000 MTMT accounts altogether, meaning that our data collection has covered around 86% of the total scientific community in the country. Figure 1B illustrates the number of all publications from the entire career of those PhD students who defended between 1993 and 2010 and from their co-authors. The data processing and analysis was carried out in R (version number 4.0.0).

Publication variables
Measuring scientific success, especially individual scientific performance is a complex problem. Traditionally, it is based on production (publication) numbers, scientific impact (citation numbers) and structural measurements for example the network characteristics of authorship (Van Balen et al., 2012; Glänzel et al., 2019). However, the raw citation number depends on several factors, such as the year of publication, research field, document type (e.g. research article, review article or proceedings), and journal characteristics (e.g. frequency of occurrence, number of articles in the journal). For example, it is easy to see that the earlier an article has appeared, the more citations it could receive. The citation habits are different in individual research fields, so to compare two citation measures we must do it in the same research area. The various document types use different number of references. Thus, the comparison is more accurate if it is made within the same document type. Moreover, the journal characteristic also can cause a bias on raw citation numbers. The solution for these problems is using normalized citation numbers.

In our case the MTMT database contains only raw citation numbers in the year when we downloaded the data, in 2020 (Figure 2 illustrates the publication and citation distributions over the examined period). To handle this problem, we compared each PhD holder in two-year periods, in a cumulative way using the year of their PhD defense as a starting

![Figure 2](image-url)
point. We compared PhD holders by their research field. The MTMT database contains document types as articles, books, others, but we were unable to distinguish research articles and review articles. Therefore, we did not consider the document types. Figure 3A shows the cumulative number of citations of the examined PhD students, while Figure 3B shows the cumulative number of their publications.

Network variables
To answer our research question, whether cohesive or diverse co-authorship network structure favors the success of a young researcher, we analyzed the weighted and dynamic ego-networks of PhD students. Such networks were generated from the publication records. These ego-networks include the PhD student in the center (ego), to which co-authors (alters in the ego-network terminology) are connected. Links are undirected but weighted by the number of co-authored papers. The networks are dynamic, such that we add new collaborators and new links to the ego-network of individual PhD students as new papers are published, but do not delete ties over the years. Since we have access to the publications of co-authors, the links between alters contain those publications that were not authored by the PhD student.

Cohesive networks are dense and include strong, high-bandwidth ties (Aral, 2016). That is, co-authors frequently publish with each other. Such network structures are thought to capture an environment, in which shared work experience and developed trust facilitate learning from peers. In cohesive networks knowledge transfer is faster and more efficient such

![Figure 3A: Cumulative number of citations by Hungarian PhD holders (year of defense 1993-2010) in 0-10 years after defense.](image)

![Figure 3B: Cumulative number of papers by Hungarian PhD holders (year of defense 1993-2010) in 0-10 years after their defense.](image)

**Figure 3. Citations and papers of PhD students.** A. Cumulative number of citations by the Hungarian PhD holders (year of defense 1993-2010) in 0-10 years after defense. B. Cumulative number of papers by Hungarian PhD holders (year of defense 1993-2010) in 0-10 years after their defense.
that the PhD student can learn complex knowledge easier (Reagens & McEvily, 2003). On the contrary, diverse networks, in which co-authors have not worked with each other but with the PhD student, capture an environment that provides the student with diverse capabilities of co-authors. In such networks, innovation and novel combination is more likely (Burt, 2001). In case the student can integrate distinct pieces of knowledge, diverse networks might help them to publish papers with high degree of novelty.

We used Burt (2000) constraint indicator that characterizes ego-networks in the cohesive-diverse continuum using the formula:

\[
\text{CON}_i = \sum_j \left( p_{ij} + \sum_q p_{iq}p_{jq} \right)^2 ; q \neq i, j, \quad (\text{Eq.1})
\]

where \( p_{ij} \) and \( p_{iq} \) is the number of papers that PhD student \( i \) has co-authored with colleagues \( j \) and \( q \), and \( p_{jq} \) is the number of papers that \( j \) and \( q \) has co-authored without \( i \). The indicator takes high values in case co-authors publish intensively together and low values are produced when co-authors do not publish together.

As the size of ego-networks grow, the probability that co-authors are connected might decrease, which has often been found in co-author networks (see for example Tóth & Lengyel, 2021). Thus, one must consider the degree of PhD students as well as their number of co-authors.

The distributions of degree and constraint are depicted in Figure 4. As expected, these two indicators change the opposite direction. The number of relations rise in time (Figure 4A) which is obvious because we used a cumulative ego network. The distribution of constraint slightly decreases (Figure 4B), and the median of constraint also falls in time. The cause is that the size of ego networks grows in time and those PhD holders who get more and more co-authors have also more diverse collaboration network.

We calculated further measures that might be also used to characterize cohesion and diversity in ego-networks. Betweenness centrality quantifies diversity in the network of PhD students by measuring the number of shortest paths in the network that go through the ego. The higher betweenness centrality of the ego the more diversity in the network. Global clustering quantifies the fraction of closed triangles in the network among all possible triangles, while network density measures the fraction of observed ties among all possible ties with the ego. The higher these measures the higher cohesion in the ego-network.

Table 1 reports Pearson correlation coefficients between network parameters at the second and eight year after PhD defense. As expected, we find a negative correlation between degree and all other network indices. Constraint is strongly correlated with network density. We have run alternative regression specifications with the network measures listed in Table 1 but only found significant results for constraint.

Regression framework
Our data enables us to capture impact of publications as a snapshot in 2020 by the total number of citations received until then. This allows for cross-sectional specification, in which we can compare students who finished in the same year and consider publications that they produced until a certain year after defense. This way, we can avoid the problem that earlier publications have more time to collect citations.

To answer the question whether cohesive co-authorship networks of PhD students during their studies help their future success, we estimate the number of accumulated citations (\( \text{CIT}_{i,t+8} \)) of student \( i \) to the paper that they published until the 8\(^{th} \) (\( t + 8 \)) year following defense at year \( t \) with the following equation:

\[
\text{CIT}_{i,t+8} = \alpha + \beta_1 \text{CIT}_{i,t+2} + \beta_2 \text{PAP}_{i,t+2} + \beta_3 \Delta \text{PAP}_{i,t+8} + \beta_4 \text{DEG}_{i,t+2} + \beta_5 \text{CON}_{i,t+2} + \theta_i + t_i + \epsilon_i, \quad (\text{Eq.2})
\]

where \( \text{CIT}_{i,t+2} \) denotes citations to papers published until the second year after defense, \( \text{PAP}_{i,t+2} \) and \( \text{PAP}_{i,t+8} \) are papers published until the second year after defense and \( \Delta \text{PAP}_{i,t+8} \) is the number of papers published between the second and eight year after defense, \( \text{DEG}_{i,t+2} \) is the degree, and \( \text{CON}_{i,t+2} \) is the constraint measure of the student’s co-author network, \( \theta_i \) is scientific area-specific fixed-effect, \( t_i \) is year dummies and \( \epsilon_i \) is the error term.
To estimate Eq. 2, we used ordinary least squares (OLS) linear regression models. The scientific area fixed effects are specified by research fields of doctoral schools. These latter refer to 54 categories of research fields defined by the Hungarian Accreditation Committee (HAC, 2018): exactly one research field has been assigned to each doctoral school. All variables are log-transformed.

Table 1. Pearson correlation of network parameters of cumulative ego networks. Two years (below diagonal) and eight years (above diagonal) after Hungarian PhD holders defense (year of defense 1993-2010).

|                  | 1   | 2  | 3   | 4   | 5   | 6   | 7   |
|------------------|-----|----|-----|-----|-----|-----|-----|
| Degree           | 1.0 | -0.26 | -0.57 | -0.44 | -0.48 | -0.42 | -0.41 |
| Betweenness centrality | -0.18 | 1.0 | 0.0 | 0.26 | 0.37 | 0.14 | 0.03 |
| Constraint       | -0.61 | -0.06 | 1.0 | 0.36 | 0.44 | 0.71 | 0.70 |
| Global clustering with ego | -0.31 | 0.12 | 0.22 | 1.0 | 0.96 | 0.46 | 0.46 |
| Global clustering without ego | -0.37 | 0.22 | 0.30 | 0.93 | 1.0 | 0.68 | 0.68 |
| Graph density with ego | -0.45 | 0.08 | 0.71 | 0.40 | 0.58 | 1.0 | 0.94 |
| Graph density without ego | -0.44 | -0.02 | 0.70 | 0.35 | 0.58 | 0.96 | 1.0 |

Figure 4. Degree and constraint of ego-networks over time. A. The distribution of degree in cumulative ego networks of the Hungarian PhD holders (year of defense 1993-2010) between 1990-2019. B. The distribution of constraint in cumulative ego networks of the Hungarian PhD holders (year of defense 1993-2010) between 1990-2019.
Results

Table 2 reports results of an OLS regression of estimating Eq.2. In columns 1-3, we estimate citations to papers that were published until the 8th year following defense with variables that capture publications and co-authorship until the 2nd year following defense. We introduce variables in a stepwise manner such that a baseline model is run in column 1 and network variables are introduced in columns 2 and 3.

Throughout the models, we found a very strong positive correlation between CIT\textsubscript{t+2} and CIT\textsubscript{t+8} that is a trivial relation but has importance in our empirical exercise. Because citations are collected for all publications in 2020, CIT\textsubscript{t+8} includes CIT\textsubscript{t+2}. However, the very high correlations also mean that most of the citations at the end of the early career stage are received to the publications that were published during or closely after PhD studies. PAP\textsubscript{t+2} is negatively correlated while ΔPAP\textsubscript{t+8} is positively correlated with the dependent variable. These findings suggest that due to accumulation of citations, the best strategy for PhD students is to produce a few but high impact papers that will help them to collect citations in their early career.

In column 2, we introduce degree that leaves correlations of other covariates almost unchanged. DEG is positively correlated with CIT\textsubscript{t+8} suggesting that the number of co-authors facilitates citations. Note that there might be various mechanisms at play; citations might grow with the number of co-authors because they can also cite the paper or spread the word, and alternatively, the project and the PhD student can gain from working with and learning from many collaborators.

Constraint is positively correlated with CIT\textsubscript{t+8}. Controlling for DEG, the number of publications, and including year and scientific field dummies, CON quantifies the extent to which co-authors of the PhD student have collaborated in publications that are published until the second year after the defense of the student. Our finding suggests that such cohesive ego-networks are beneficial for PhD students. Because we also control for the citations to papers, this finding confirms that PhD students benefit the most from working in cohesive collaboration networks because these create efficient learning environments.

Correlations of independent variables indicate that the models are not violated by multicollinearity. The highest value of the Pearson correlation coefficients is $\rho = 0.41$ between DEG and PAP\textsubscript{t+2}. We document the correlation between DEG and CON in Table 1 ($\rho = 0.61$), but the inclusion of these variables together is conceptually motivated as we describe before. Further, the inclusion of CON in Model 3 does not substantially influence the coefficient of DEG.

In Table 3, we report full regression models decomposed into four big scientific areas such as Science, Life Science, Engineering, Social Science. To achieve these scientific areas, we have grouped the 54 scientific fields (as per National Accreditation Committee). We found that DEG and CON is significant for the Life Science subsample while DEG only is

|                  | (1)          | (2)          | (3)          |
|------------------|--------------|--------------|--------------|
| CIT (log)        | 0.899***     | 0.894***     | 0.893***     |
|                  | (0.009)      | (0.009)      | (0.009)      |
| PAP (log)        | −0.211***    | −0.217***    | −0.226***    |
|                  | (0.017)      | (0.019)      | (0.020)      |
| Δ PAP (log)      | 0.317***     | 0.312***     | 0.313***     |
|                  | (0.010)      | (0.011)      | (0.011)      |
| DEG (log)        | 0.043**      | 0.087***     |              |
|                  | (0.018)      | (0.028)      |              |
| CON (log)        |              | 0.244**      |              |
|                  |              | (0.121)      |              |
| Constant         | 2.327***     | 2.324***     | 2.141***     |
|                  | (0.374)      | (0.372)      | (0.383)      |
| N                | 2,061        | 1,948        | 1,948        |
| R2               | 0.919        | 0.917        | 0.918        |

Note: *p < 0.1; **p < 0.05; ***p < 0.01.
weakly significant in Engineering. Thus, cohesive research environment is important for Life Science students and less for students in other fields.

**Discussion**

In this study we examined the success of students who defended theses in Hungarian doctoral schools between 1990 and 2010 by looking at their publication records and accumulated citations in 2019. Our bibliometric database also contains the PhD students’ publications and their co-authors’ publications between 1990 and 2019. We analyzed whether cohesive or diverse co-author network structure gives a better chance to a young researcher to stand out in terms of citations eight years after defense. Linear regression models suggest that those students who participate in cohesive collaboration networks, receive significantly more citations at the end of their career. This result highlights the need for strong collaborations and effective learning environment during doctoral studies. However, our results regarding the structure of co-author networks are specific to Life Science students. Thus, cohesion is mostly important in areas where new knowledge is produced in teamwork.

The present paper contributes to a growing literature, in which studies try to determine factors that support the future success of young researchers. Li and co-authors (2019) demonstrate that those students who publish with top scientists had a greater chance of being more successful 20 years later. Moreover, this effect is more important in the case of PhD students affiliated with a less prestigious PhD school. Sarigöl et al. (2014) illustrate a similar phenomenon: a paper gets more citations if its’ authors are central in the large co-author network of their field. We add to this discussion by studying the co-author ego-networks of PhD students. Our findings confirm that the structure of the group collaboration matters for the future academic career of students.

We also find that those students are more successful, measured in citations, who focus on few papers. These results are robust across all large scientific fields. By concentrative efforts into a small number of publications, students can achieve higher quality papers that might be accepted to better journals. Because citations typically demand several years to accumulate, students need high-impact papers at the beginning of their career to stand out later when they are at the end of the early-career stage. This can help them in research proposals and thus facilitate academic careers in the long run.

**Data availability**

**Underlying data**

Raw data, as accessed through the Hungarian Scientific Bibliography Database, are not publicly available due to privacy considerations of the authors. Access can be requested through the Library and Information Centre of the Hungarian Academy of Sciences; contact details can be found on the library’s website https://www.mtmt.hu/kapcsolat. Access to the data will be provided under the following conditions: the researcher presents legitimate research purposes to the ad-hoc

| Table 3. Estimates for citations eight years after defense by scientific areas. OLS regressions with year and scientific field fixed effects and robust standard errors. |
|---------------------------------------------------------------|
| Sciences | Life Sciences | Engineering | Social Sciences |
| CIT_{t2} (log) | 0.919*** (0.020) | 0.870*** (0.017) | 0.917*** (0.034) | 0.910*** (0.026) |
| PAP (log) | –0.294*** (0.046) | –0.220*** (0.033) | –0.222*** (0.072) | –0.267*** (0.058) |
| Δ PAP (log) | 0.357*** (0.023) | 0.296*** (0.018) | 0.255*** (0.034) | 0.388*** (0.033) |
| DEG (log) | 0.068 (0.063) | 0.162** (0.054) | 0.191* (0.110) | 0.039 (0.075) |
| CON (log) | 0.173 (0.255) | 0.697*** (0.274) | 0.654 (0.401) | 0.060 (0.303) |
| Constant | 2.149*** (0.438) | 2.324*** (0.372) | 1.124*** (0.422) | 0.434 (0.536) |
| N | 437 | 1,948 | 155 | 279 |
| R2 | 0.919 | 0.917 | 0.942 | 0.910 |

Note: *p < 0.1; **p < 0.05; ***p < 0.01.
The authors acknowledge the kind help of Gusztáv Ladányi in collecting the bibliometric data and Dániel Horváth in matching the bibliometric data with thesis defense data.

Zenodo: Data and Code for the manuscript PhD students in life sciences can benefit from team cohesion, https://doi.org/10.5281/zenodo.5129288 (Vida et al., 2021).

This project contains the following underlying data:

- Phd_data.csv (contains anonymized author ID, network and publication variables after 2–4–6–8 years of PhD defense, and discipline of PhD thesis).

- Phd_regressions_code.R (analysis code used in the study).

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements

The ethical committee of the Library and Information Centre of the Hungarian Academy of Sciences and after the committee’s positive decision, signs a non-disclosure agreement.

References

Aral S: The future of weak ties. *Am J Sociol*. 2016; 121(6): 1931–1939. Published Full Text

Aral S, Van Alstyne M: The diversity-bandwidth trade-off. *Am J Sociol*. 2011; 117(1): 90–171. Published Full Text

Beaver DD: Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*. 2001; 52(3): 365–377. Published Full Text

Boh T, de Vaan M, van de Rijt A: The Matthew effect in science funding. *Proc Natl Acad Sci U S A*. 2018; 115(19): 4887–4890. Published Abstract | Publisher Full Text | Free Full Text

Burt RS: Structural holes: The social structure of competition. Harvard University Press; 2009.

Burt RS: The network structure of social capital. *Res Organizational Behav*. 2000; 22: 345–423. Published Full Text

Clauser A, Larremore DB, Sinatra R: Data-driven predictions in the science of science. *Science*. 2017; 355(6324): 477–480. Published Abstract | Publisher Full Text

De Vaan M, Stark D, Vedres B: Game changer: The topology of creativity. *Am J Sociol*. 2015; 120(6): 1144–1194. Published Full Text

de Solla Price DJ, Beaver D: Collaboration in an invisible college. *Am Psychol.* 1966; 21(11): 1011–1018. Published Abstract | Publisher Full Text

Fortunato S, Bergstrom CT, Bornkamm G, et al.: Science of science. *Science*. 2018; 359(6379), eaao0185. Published Abstract | Publisher Full Text

Glänzel W, Schubert A: Analysing scientific networks through co-authorship. In: Moed HF, Glänzel W, Schmoch U (eds.), *Handbook of quantitative science and technology research*. Dordrecht: Springer; 2004; (pp. 257–270).

Hood W, Wilson C: The literature of bibliometrics, scientometrics and informetrics. *Scientometrics*. 2001; 52(2): 291–314. Published Full Text

Hungarian Accreditation Committee: Doctoral branches of sciences. 2018. Reference Source

Juhász S, Tóth G, Lengyel B: Brokering the core and the periphery: Creative success and collaboration networks in the film industry. *Plos ONE*. 2020; 15(2): e0229436. Published Abstract | Publisher Full Text

Katz JS, Martin BR: What is research collaboration? *Res policy*. 1997; 26(1): 1–18. Published Full Text

Li W, Aste T, Caccioli F, et al.: Early coauthorship with top scientists predicts success in academic careers. *Nat Commun*. 2019; 10(1): 1–9. Published Abstract | Publisher Full Text | Free Full Text

Luukkonen T, Tijssen R, Persson O, et al.: The measurement of international scientific collaboration. *Scientometrics*. 1993; 28(1): 15–36. Published Full Text

Ma Y, Mukherjee S, Uzzi B: Mentorship and protégé success in STEM fields. *Proc Natl Acad Sci U S A*. 2020; 117(25): 14077–14083. Published Abstract | Publisher Full Text | Free Full Text

Mali F, Kranegger L, Doreian P, et al.: Dynamic scientific co-authorship networks. In: Schanenhorst A, Birner K, van den Besselaar P (eds.) *Models of science dynamics*. Berlin, Heidelberg: Springer; 2012; (pp. 195–232).

Mukherjee S, Huang Y, Neidhardt J, et al.: Prior shared success predicts victory in team competitions. *Nat Hum Behav*. 2019; 3(1): 74–81. Published Abstract | Publisher Full Text | Free Full Text

Melin G, Persson O: Studying research collaboration using co-authorships. *Scientometrics*. 1996; 36(3): 363–377. Published Full Text

Reagans R, McEvily B: Network structure and knowledge transfer: The effects of cohesion and range. *Admin Sci Quarterly*. 2003; 48(2): 240–267. Published Full Text

Sarigöl E, Pfitzner R, Scholtes I, et al.: Predicting scientific success based on coauthorship networks. *EPJ Data Sci*. 2014; 3: 1–16. Published Full Text

Sonnenwald DH: Scientific collaboration. In: Cronin B (szerk.), *Annual Review of Information Science and Technology*. Medford (NJ): Information Today; 2007; Vol. 41 (pp. 643–681).

Tortoriello M, Reagans R, McEvily B: Bridging the knowledge gap: The influence of strong ties, network cohesion, and network range on the transfer of knowledge between organizational units. *Organization Sci.* 2012; 23(4): 1024–1039.
Tóth G, Lengyel B: Inter-firm inventor mobility and the role of co-inventor networks in producing high-impact innovation. J Technology Transfer. 2021; 46: 117–137. Publisher Full Text

Van Balen B, Van Arensbergen P, Van Der Weijden I, et al.: Determinants of success in academic careers. Higher Education Policy. 2012; 25(3): 313–334. Publisher Full Text

Vedres B: Forbidden triads and creative success in jazz: the Miles Davis factor. Appl Netw Sci. 2017; 2(1): 1–25. PubMed Abstract | Publisher Full Text | Free Full Text

Vida Z, Járay I, Lengyel B: Data and Code for the manuscript PhD students in life sciences can benefit from team cohesion [Data set]. Zenodo. 2021. Publisher Full Text

Waltman L, Calero-Medina C, Kosten J, et al.: The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. J Am Society Information Sci Technol. 2012; 63(12): 2419–2432. Publisher Full Text

Wang Y, Jones BF, Wang D: Early-career setback and future career impact. Nat Commun. 2019; 10(1): 1–10. PubMed Abstract | Publisher Full Text | Free Full Text

Wuchty S, Jones BF, Uzzi B: The increasing dominance of teams in production of knowledge. Science. 2007; 316(5827): 1036–1039. PubMed Abstract | Publisher Full Text

Zeng A, Fan Y, Di Z, et al.: Fresh teams are associated with original and multidisciplinary research. Nat Hum Behav. 2021. PubMed Abstract | Publisher Full Text

Ziman J: Prometheus Bound Science in a Dynamic Steady State. Cambridge: Cambridge Univ. Press; 1994.
The paper presents an interesting study on team cohesion and its effects on PhD students’ success based on a novel dataset. While the paper is well-written, I list suggestions for further development and three textual issues.

1. **Embedding in the literature:** There is a sizeable literature by now on the (collaborative) conditions that support PhD students. I list below some references. I do not suggest that you have to integrate all these studies in your paper, but rather that you read the papers and see which ones are relevant to your study. Taking some of these papers into account would better embed your study in the current literature and also better highlight what your contribution is.

2. **Limitations:** You explain well your data and methodology but some of the limitations are not well spelled out. First, you look only at PhD students who finished by writing the thesis. Hence, the ones who did not finish, arguably the least successful ones, are left out. This means that - in a way - you “sample on the dependent variable” in the sense that your dependent variable is a success variable, while you analyse only those who successfully defended the thesis. I do not see this as a major issue, but I think it should be highlighted as a limitation.

Second, on page 4 you report that 60 percent of the PhD students could be identified in the MTMT dataset. Here, I expected more information why the remaining 40 percent could not be matched. I understand from page 5 that the co-author network of a PhD student that is collected, comprises only of authors in MTMT, which I reckon excludes foreign co-authors. Please clarify. And, if my comment is correct, you could also highlight this as a data limitation.

Third, it is further also unclear whether the “raw citation numbers” in the MTMT database include only citations by authors in the MTMT. Again, please clarify. And, if my comment is correct, you could also highlight this as a data limitation.
3. **Results**: page 8, table 1: why do you report a correlation matrix including multiple variables that are NOT included in the regression analysis later on? Rather uncommon. Instead, I suggest you show first a table with the descriptive statistics only of the variables included in the regression analysis and then also a correlation matrix, again, only for the variables included in the regression analysis later on.

page 9, table 2: You do OLS regressions but your dependent variable is a count variable which is most probably also skewed. Please look into this again. Most probably, you have to do Poisson or Negbin regressions instead. (see e.g.: Cameron, C., Trivedi, P.K., 1998. Regression analysis of count data. Cambridge University Press, New York.)

4. **Text**: page 3: “but the impact of these papers correlates strongly with future impact” -> I do not understand this sentence

page 5: what do you mean by “raw”

page 10 “its’” should be “its”

**References**

1. Baruffaldi S, Visentin F, Conti A: The productivity of science & engineering PhD students hired from supervisors' networks. *Research Policy*. 2016; 45 (4): 785-796 [Publisher Full Text]
2. Broström A: Academic breeding grounds: Home department conditions and early career performance of academic researchers. *Research Policy*. 2019; 48 (7): 1647-1665 [Publisher Full Text]
3. Chariker JH, Zhang Y, Pani JR, Rouchka EC: Identification of successful mentoring communities using network-based analysis of mentor-mentee relationships across Nobel laureates. *Scientometrics*. 2017; 111 (3): 1733-1749 [PubMed Abstract | Publisher Full Text]
4. Dericks G, Thompson E, Roberts M, Phua F: Determinants of PhD student satisfaction: the roles of supervisor, department, and peer qualities. *Assessment & Evaluation in Higher Education*. 2019; 44 (7): 1053-1068 [Publisher Full Text]
5. Feldon D, Litson K, Jeong S, Blaney J, et al.: Postdocs’ lab engagement predicts trajectories of PhD students’ skill development. *Proceedings of the National Academy of Sciences*. 2019; 116 (42): 20910-20916 [Publisher Full Text]
6. Heinisch D, Buenstorf G: The next generation (plus one): an analysis of doctoral students’ academic fecundity based on a novel approach to advisor identification. *Scientometrics*. 2018; 117 (1): 351-380 [Publisher Full Text]
7. Horta H, Santos J: The Impact of Publishing During PhD Studies on Career Research Publication, Visibility, and Collaborations. *Research in Higher Education*. 2016; 57 (1): 28-50 [Publisher Full Text]
8. Liu J, Tang T, Kong X, Tolba A, et al.: Understanding the advisor–advisee relationship via scholarly data analysis. *Scientometrics*. 2018; 116 (1): 161-180 [Publisher Full Text]
9. Long J, McGinnis R: The effects of the mentor on the academic career. *Scientometrics*. 1985; 7 (3-6): 255-280 [Publisher Full Text]
10. Malmgren RD, Ottino JM, Nunes Amaral LA: The role of mentorship in protégé performance. *Nature*. 2010; 465 (7298): 622-6 [PubMed Abstract | Publisher Full Text]
11. Paglis L, Green S, Bauer T: Does adviser mentoring add value? A longitudinal study of mentoring and doctoral student outcomes. *Research in Higher Education*. 2006; 47 (4): 451-476 [Publisher Full Text]
12. Rossi L, Freire I, Mena-Chalco J: Genealogical index: A metric to analyze advisor–advisee
Is the work clearly and accurately presented and does it cite the current literature?  
Partly

Is the study design appropriate and is the work technically sound?  
Partly

Are sufficient details of methods and analysis provided to allow replication by others?  
Partly

If applicable, is the statistical analysis and its interpretation appropriate?  
Partly

Are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions drawn adequately supported by the results?  
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** scientometrics, innovation studies

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com