The Formation of Co-authorship Networks - A Case Study of the SOCIUM Research Institute

Christine Hedde - von Westernhagen*

05.08.2022

*c.hedde-vonwesternhagen@students.uu.nl

Introduction

Over the course of the last century - and more rapidly so for the past twenty years - the amount of co-authored scientific literature has steadily increased in nearly every discipline (Henriksen 2016). This development has also been emergent in the social sciences where co-authorship has traditionally been much less prominent than in the natural sciences (e.g. Wallace, Larivière, and Gingras 2012; Wuchty, Jones, and Uzzi 2007). When investigating co-authorship behavior, research most often draws on large bibliographic databases which are mainly comprised of peer-reviewed journal articles. In the social sciences, however, other publication types like books and book chapters, as well as national and non-scholarly publications, still make up a large share of scientific output (Hicks 2005). The field is therefore relatively understudied with regards to assessing empirical collaboration behavior (for an exception see Leifeld 2018).

This study aims at counteracting the neglect of social science co-authorship practices in the literature by accessing the complete scientific publication data of a well-established research center at the University of Bremen. The dataset encompasses a network of 240 authors over the course of ten years (2010-2019). Alongside the bibliographic entries, several other features to investigate the generative process behind this case of co-authorship were encoded during the data collection process. To account for the interdependent as well as temporal nature of the data, I employ the temporal version of an exponential random graph model (TERGM) to explain the emergence of co-authorship, i.e., the formation of ties in this network.

I proceed as follows: The next section provides a theoretical foundation of scientific collaboration practices, highlighting differences between the social sciences and other disciplines. Section three gives a brief introduction to the case of the SOCIUM research center. In the fourth section, I describe the data and statistical methodology used to investigate co-authorship emergence. Results are presented in section five, followed by a conclusion.

Collaboration practices in the social sciences

In defining collaborative research, the broadest perspective would consider even a helpful remark at a conference as part of the process, and advancing science would be an endeavor of the entire global scientific community. Far more commonly, however, scientometric studies are conducted that use co-authorship or citation behavior as indicators for collaboration. Only few examples exist that combine qualitative and quantitative approaches (e.g. Laband and Tollison 2000). Another distinction has to be made between inter- and intra-disciplinary research. Even though collaboration has generally increased for both types, there exist noteworthy particularities for the specific disciplines (van Rijnsoever and Hessels 2011).

When looking at driving forces of collaborative research, an increased resource intensity can be attributed to both the natural and social sciences. While for the natural sciences material aspects like laboratories, technical supply, and the respective funding play a large role (Cronin 2001; Wuchty, Jones, and Uzzi 2007),

the need of more diverse knowledge resources in the form of experts for quantitative methods and statistics is of more importance in the social sciences (Fisher et al. 1998; Moody 2004).

Another general development in scientific practice that incentivizes co-authorship has been the growing importance of publication for career advancement, especially for young researchers (Hangel and Schmidt-Pfister 2017). This tendency is reenforced by the establishment of performance measures like impact factors or citation indices for the evaluation of researchers (Ingwersen and Larsen 2014; Ossenblok, Engels, and Sivertsen 2012). Empirical evidence suggests that collaboration is indeed highly beneficial for researchers in terms of research impact and career advancement (Li, Liao, and Yen 2013; Lutter and Schröder 2016; Wuchty, Jones, and Uzzi 2007).

Acknowledging this positive trend in collaborative practices, the question arises how researchers choose their co-authors. Two of the most important contributing factors for co-authorship are likely to be topic similarity and publication strategy (e.g. Leifeld 2018). Within research domains, however, it has become more popular to collaborate with people having complementary skills, most often specialized analysis methods (Moody 2004). Another empirical pattern is that of supervisory relationships, where senior researchers co-author the work of multiple graduate students. Collaboration between multiple senior researchers, on the other hand, is less common. While the trend has been stagnating over the past years, geographic location also still plays a role in the choice of collaboration partners (Hunter and Leahey 2008). The prevalence of same- or cross-gender collaboration seems to be largely dependent on the gender composition of the discipline; the same holds for the general proneness of the respective gender to collaborate (Hunter and Leahey 2008; Laband and Tollison 2000).

The SOCIUM research center

In order to assess the circumstances under which the results of this study apply and in how far they might be generalizable, I lay out the characteristics of the case at hand. The SOCIUM Research Center on Inequality and Social Policy (SOCIUM)

> 'is the only German research institute in social sciences which deals with theoretical and empirical questions of inequality, social policies and their social and political interdependencies. The interdisciplinary research focuses on the social, economic, political, cultural, organizational, legal, historical and sociomedical conditions and effects of social inequality, public social policies and their interdependencies.' (Universität Bremen 2022)

It currently has 153 active members, most of whom have a background in sociology or political science, but also scientists from related disciplines are present. They together approach the aforementioned topics from within six thematic departments:

1. Theoretical and Normative Foundations

2. Political Economy of the Welfare State
3. Dynamics of Inequality in Welfare Societies
4. Life Course, Life Course Policy, and Social Integration
5. Health, Long-Term Care and Pensions
6. Methods Research

Especially the health department differs disciplinarily, having members with a background in pharmacology, epidemiology, public health and administration. All the departments have several working groups that take a narrower thematic approach. Within these working groups researchers work on specific projects that are mostly publicly funded.

The SOCIUM is part of the University of Bremen, a medium-sized public university in Germany, and the SOCIUM is not the only social scientific research institute at the university. In the same building also the Research Centre for International Relations, European Politics, and Political Theory (InIIS) has its offices, and both institutes are part of the Collaborative Research Centre 1342 "Global Dynamics of Social Policy" (CRC). Some members of the CRC that are not part of the SOCIUM or InIIS are located in the mentioned building as well. Additionally, the Bremen International Graduate School of Social Sciences (BIGSSS) is a cooperating partner of the SOCIUM, and a part of the graduate students' offices are also in that same building. Naturally, members collaborate across those institutes, and an analysis covering all institutes might paint an even more realistic picture.

Data and Methods

Data on co-authorship

A network of co-authorship relations has been constructed from the publication bibliographies that are available on the SOCIUM's website[1]. Included are monographs, edited volumes, journal articles, articles in edited volumes, working and discussion papers, and also grey literature, overall making this study more representative of the publication behavior in the social sciences (Hicks 2005). In total, this results in 2278 publications by 1239 authors, 240 of which are or have been members or affiliated members of the institute (as listed on the website). The data was collected for a time span of ten years (January 2010 until December 2019) and is thus available for analysis in ten annual snapshots.

Alongside the bibliographic entries, the dataset has been manually augmented by multiple individual author attributes[2]: gender, department affiliation, being head of a department or a working group, and the total number of publications. Even though there are formally six departments, membership in the methods department only occurs as a secondary membership. For analysis purposes, department affiliation has

[1] www.socium.uni-bremen.de/veroeffentlichungen/
[2] All attributes were coded only for the subset of 240 SOCIUM members.

therefore been reduced to the first five departments.

Statistical methods

The goal of the analysis lies in explaining the emergence of the co-authorship network as it is observed in ten annual snapshots of the SOCIUM. A suitable tool that explicitly models endogenous network dependencies as well as exogenous factors is the exponential random graph model (ERGM), which can be extended to accommodate temporal dependencies as introduced by Hanneke, Fu, and Xing (2010). The T(emporal)ERGM can be expressed as

$$P(N^t|N^{t-K},...,N^{t-1},\theta) = \frac{e^{\theta^\top \Gamma(N^t,N^{t-1},...,N^{t-K})}}{\sum_{i=1}^{\mathbb{N}} e^{\theta^\top (N_i^t,N_i^{t-1},...,N_i^{t-K})}}$$

which represents the probability P of observing a given network at a given time point conditional on K previously observed networks, but it can equally be interpreted as the probability of a tie between each dyad conditional on the rest of the network. The parameter vector θ contains the model coefficients that have to be estimated and Γ is a vector of explanatory variables consisting of endogenous network statistics as well as exogenous variables computed at each time point. The denominator is the sum over the set of all possible permutations $\mathbb{N}$ of the network with the same number of nodes and serves as a normalizing constant. In the following, I introduce the explanatory variables used in this study.

Endogenous model terms

The edges term simply sums the number of edges in a specific network. It serves as a baseline variable analogous to a regression intercept and is defined as

$$\Gamma_{edges}(N^t) = \sum_{i \neq j} N_{ij}^t.$$

Resulting from the observation that researchers are increasingly prone to distribute workload to co-authors with complementary skills, I introduce a model term that captures the tendency of two co-authors having multiple other shared co-authors, in turn manifesting as triadic closure. This term for edge-wise shared partners is further modified by a geometrical decay parameter α that captures how having one or two shared partners is more frequent than, e.g., five or six:

$$\Gamma_{gwesp}(N^t,\alpha) = e^{\alpha} \sum_{h=1}^{n-2} (1-(1-e^{-\alpha})^h) ESP_h(N^t),$$

where $ESP_h(N^t)$ is the number of edges with h shared partners. In attempting to optimize model fit, the parameter α is set to 0.3 in the following analysis. Similarly, the number of co-authors researchers have across

the network follows an exponentially decaying distribution. This can be attributed to the higher output of relatively fewer senior researchers compared to many junior researchers who have just started publishing. The corresponding term for the geometrically weighted degree distribution is defined as

$$\Gamma_{gwd}(N^t, \lambda) = e^{\lambda} \sum_{h=1}^{n-1} (1 - (1 - e^{-\lambda})^h) D_h(N^t),$$

where $D_h(N^t)$ is the number of nodes having a degree centrality of h, and decay parameter $\lambda = 0.4$. Considering the temporal nature of the data, I test for the extent of edge formation from one year to the next by using the edge innovation term

$$\Gamma_{innovation}(N^t, N^{t-1}) = \sum_{i \neq j} N_{ij}^t (1 - N_{ij}^{t-1})$$

which counts the number of edges created at t that were not present in $t-1$. As co-authorship is a relationship that (usually) cannot be dissolved once established, a term to test for stability or loss of edges would not be a meaningful addition to this model.

Exogenous model terms

Based on the assumption of possibly different collaboration behaviors I add two node covariate terms to the model, one for department affiliation and one for gender, which are computed as

$$\Gamma_{nodecov}(N^t, \mathbf{x}) = \sum_{i \neq j} x_i N_{ij},$$

where $\mathbf{x}$ is a vector of covariate values for all nodes. Furthermore, I assume certain homophiliy in co-authorship choice. People in the same department might be more prone to collaborate since their topics are more similar and they often work on the same projects. I also check for gender homophily. The node-match term

$$\Gamma_{nodematch}(N^t, \mathbf{x}) = \sum_{i \neq j} [x_i = x_j] N_{ij}$$

adds one to the count when two adjacent nodes have the same covariate value, i.e., when $x_i = x_j$ (and 0 otherwise). A term counting not equal but unequal covariate values is the node-mix term. I use it to test for the prevalence of supervisor-supervisee relationships by evaluating whether department or working group leaders rather collaborate with other status groups than amongst each other. This does not fully capture the hierarchical structure of the SOCIUM but should pick up a reasonable amount of variation in co-authorship behavior since the department and group heads encompass a large share of the senior researchers in the institute.

$$\Gamma_{nodemix}(N^t, \mathbf{x}) = \sum_{i \neq j} [x_i \neq x_j] N_{ij}.$$

Estimation procedure

The standard routine for ERGMs uses Maximum Likelihood parameter estimation based on Markov Chain Monte Carlo sampling (MCMC-MLE). This procedure, however, becomes highly computationally intensive with increasing network size and an increasing number of networks in the temporal case. It is furthermore prone to result in degeneracy when the model is not accurately enough reflecting the data generating process. I therefore resort to maximum pseudo-likelihood estimation (MPLE) which has shown to be a consistent estimator with increasing network size and number of time points (Strauss and Ikeda 1990) and does not suffer from degeneracy issues. To correct for bias in standard errors yielded by this procedure I employ a nonparametric bootstrap as proposed by Desmarais and Cranmer (2012).

Results

Figure 1 gives an impression of the development of the SOCIUM co-authorship network over time. While initially there was only a minimal amount of collaboration, by the end of 2019 the network has become increasingly connected. Results of the TERGM estimation are presented in Table 1. It can be seen that many of the included model terms significantly contribute to the emergence of the observed network, since their estimates' confidence intervals do not contain zero. All interpretations of the coefficients are conditional on the other current processes in the network as well as the networks from previous time points.

The coefficients of the endogenous terms for edge-wise shared partners and the degree distribution are both significant. This means that while the network exhibits more edge-wise shared partners as expected in a random network, the coefficient for degree is negative, implying a lower average degree as expected.

Looking at the department affiliation, department three and four differ significantly from department one (the reference category), both being more likely to show co-authorship behavior. There seems to be no difference for department two and five compared to the reference category. The significant coefficient for matching department affiliation indicates that the network exhibits more homophily regarding this factor as expected under randomness. Converting the coefficient to an odds ratio, it can be concluded that co-authorship is 7.5 times more likely to occur for researchers from the same department.

Both gender coefficients being insignificant leads to the interpretation that it is neither more likely for one specific gender to collaborate more than the other nor is there a preference to collaborate with the same gender. There is no evidence either for increased mixed-status collaboration (the reference category) compared to two department/working group leaders collaborating. Mixed-status collaboration is, however, more likely to occur than two non-leaders collaborating by a factor of ~4.

Table 1: Results of the temporal exponential random graph model using maximum pseudo-likelihood estimation and bootstrapped 95 percent confidence intervals. Decay parameters: GWESP 0.3, GWD 0.4. Refrence categories of categorial variables: Department 1, males, mix non-lead and lead.

	TERGM results
Edges	17.010
	[16.186, 28.599]
GWESP	1.854
	[1.662, 2.062]
GWD	−0.514
	[−0.923, −0.189]
Department 2	−0.081
	[−0.369, 0.252]
Department 3	0.529
	[0.118, 0.885]
Department 4	0.575
	[0.070, 1.003]
Department 5	0.349
	[−0.020, 0.713]
Match Department	2.017
	[1.698, 2.444]
Female	−0.002
	[−0.190, 0.180]
Match Female	0.053
	[−0.103, 0.202]
Mix non-lead and non-lead	1.384
	[0.970, 1.817]
Mix lead and lead	1.396
	[−13.193, 2.519]
Edge innovation	−26.418
	[−38.167, −25.057]

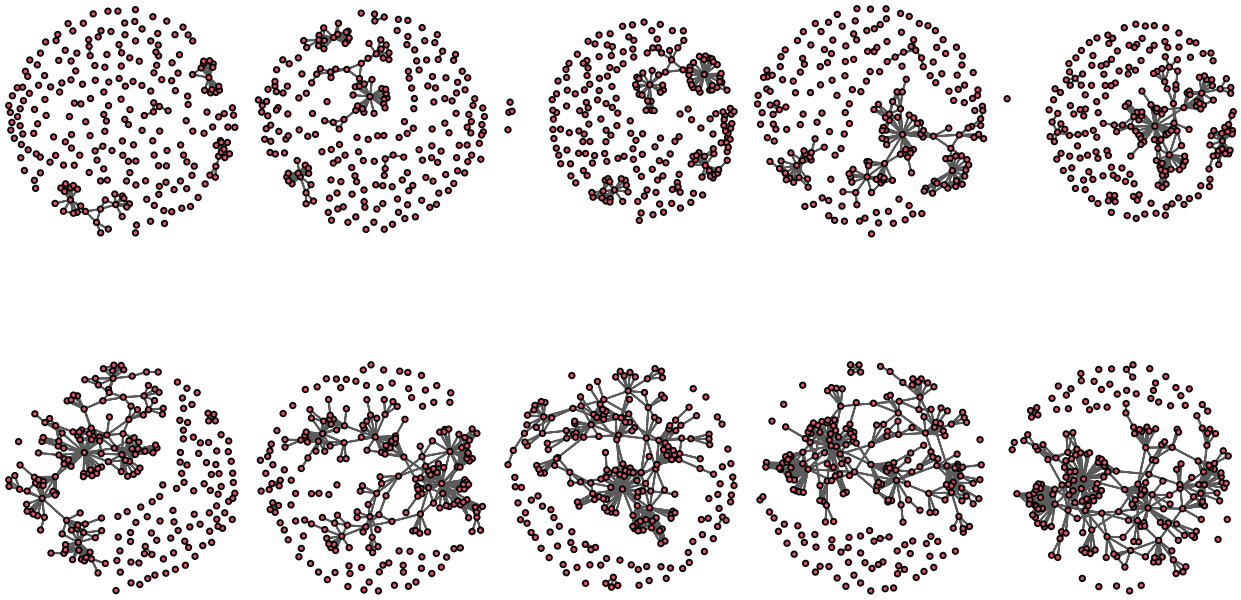

Figure 1: Evolution of the socium co-authorship network in yearly snapshots from 2010 to 2019 (top left to bottom right).

The large and significant negative coefficient of the temporal edge innovation term translates to a decreased likelihood of edge formation between nodes in t that were not connected in $t-1$ as compared to a random network. This also becomes evident in the substantive amount of isolates that remain until the last time point (see Figure 1) which would likely not occur under random edge formation.

Goodness of fit

In terms of endogenous fit, the model seems to reflect the data generating process very well. This can be seen from the first five plots in Figure 2: For each statistic, the solid black line depicting the observed network statistic aligns well with the medians of the underlying boxplots. The boxplots are based on 100 networks that were simulated using the estimated model parameters.

Furthermore, the last plot in Figure 2 assesses how well the exact placement of the edges can be predicted from the model. This is of interest since the preceding endogenous fit measures do not consider where certain edges are placed but only the amount and structures in which they appear. It is apparent from both the precision-recall curve (dark blue) as well as the receiver-operating characteristics curve (dark red) that the model predicts the location of edges extremely well, since both curves have an overall large distance to the respective light shaded curves representing the null model.

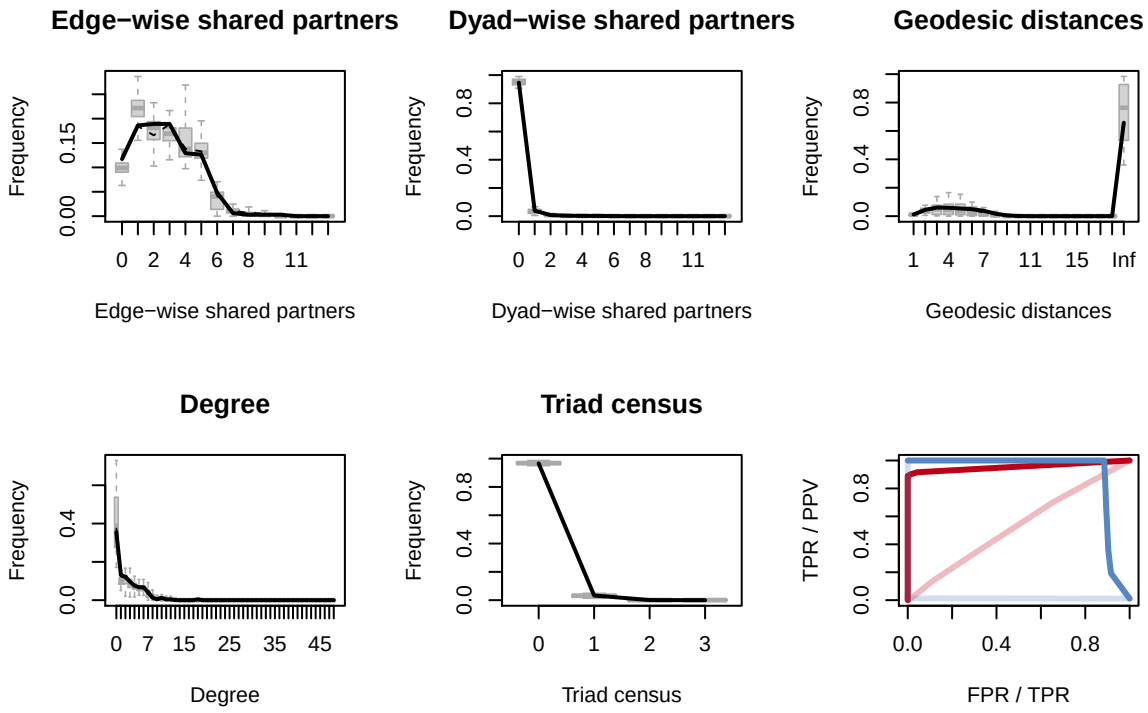

Figure 2: Goodness-of-fit plots. Bottom right presents ROC (red) and precision-recall curves (blue), while the other measures assess endogenous model fit.

Conlcusion

In this study, I investigated the emergence of co-authorship for the case of the SOCIUM research center. I drew on a bibliographic dataset that is highly representative for publication behavior in the social sciences, as it contains not only journal articles but also all other forms of scientific output. For the analysis, I employed a temporal ERGM that explicitly models the development of the network in annual intervals and its endogenous dependencies.

The results are in line with previous findings in several aspects: The amount of triadic closure as captured by edge-wise shared partners is higher than expected under randomness, while the degree distribution is skewed, having a lower average than expected (Leifeld 2018; Moody 2004). Furthermore, department membership as well as department homophily play an important role in co-authorship emergence. This can be viewed as support for findings that identify publication strategy and topic similarity as important factors (Leifeld 2018). I do not find evidence for gender differences or preferences in collaboration behavior. There is partial support in the analysis for mixed-status co-authorship being more likely than between researchers with the same institutional status. Lastly, even though the network becomes more connected over the years, there are still many authors that remain without any publications co-authored with other SOCIUM members, which is reflected in a negative edge innovation coefficient.

There are several caveats I would like to highlight, which could be improved upon in future work: First of all, it has to be kept in mind that this is a case study, and thus the extent of generalizability is debatable. Even though the SOCIUM researchers are an extensive cross-section of social science disciplines, the institute's main focus is social inequality. Furthermore, the health department is a bit of an outlier topic-wise. Also, there is a large share of extra-institutional collaboration, which is not captured in the subset of the data used in the analysis. This likely influenced the negative coefficient of edge innovation. The question remains, however, where to draw the boundary of the network.

Future work could furthermore extend the dataset, firstly, temporally, by collecting more recent as well as older publication data of the institute. And secondly, by adding more exogenous attributes, which would allow a more fine-grained analysis. Examples are the type of publication, citation indices of the authors, academic titles and positions, language of the publication, and the exact entry and exit of a member to the institute.

References

Cronin, Blaise. 2001. "Hyperauthorship: A Postmodern Perversion or Evidence of a Structural Shift in Scholarly Communication Practices?" *Journal of the American Society for Information Science and Technology* 52 (7): 558–69. https://doi.org/10.1002/asi.1097.

Desmarais, B. A., and S. J. Cranmer. 2012. "Statistical Mechanics of Networks: Estimation and Uncertainty." *Physica A: Statistical Mechanics and Its Applications* 391 (4): 1865–76. https://doi.org/10.1016/j.physa.2011.10.018.

Fisher, Bonnie S., Craig T. Cobane, Thomas M. Vander Ven, and Francis T. Cullen. 1998. "How Many Authors Does It Take to Publish an Article? Trends and Patterns in Political Science." *PS: Political Science and Politics* 31 (4): 11.

Hangel, Nora, and Diana Schmidt-Pfister. 2017. "Why Do You Publish? On the Tensions Between Generating Scientific Knowledge and Publication Pressure." *Aslib Journal of Information Management* 69 (5): 529–44. https://doi.org/10.1108/AJIM-01-2017-0019.

Hanneke, Steve, Wenjie Fu, and Eric P. Xing. 2010. "Discrete Temporal Models of Social Networks." *Electronic Journal of Statistics* 4 (none). https://doi.org/10.1214/09-EJS548.

Henriksen, Dorte. 2016. "The Rise in Co-Authorship in the Social Sciences (1980–2013)." *Scientometrics* 107 (2): 455–76. https://doi.org/10.1007/s11192-016-1849-x.

Hicks, Diana. 2005. "The Four Literatures of Social Science." In *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, edited by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, 473–96. Dordrecht: Springer Netherlands.

Hunter, Laura, and Erin Leahey. 2008. "Collaborative Research in Sociology: Trends and Contributing Factors." *The American Sociologist* 39 (4): 290–306. https://doi.org/10.1007/s12108-008-9042-1.

Ingwersen, Peter, and Birger Larsen. 2014. "Influence of a Performance Indicator on Danish Research Production and Citation Impact 2000–12." *Scientometrics* 101 (2): 1325–44. https://doi.org/10.1007/s11192-014-1291-x.

Laband, David N., and Robert D. Tollison. 2000. "Intellectual Collaboration." *Journal of Political Economy* 108 (3): 632–62. https://doi.org/10.1086/262132.

Leifeld, Philip. 2018. "Polarization in the Social Sciences: Assortative Mixing in Social Science Collaboration Networks Is Resilient to Interventions." *Physica A: Statistical Mechanics and Its Applications* 507 (October): 510–23. https://doi.org/10.1016/j.physa.2018.05.109.

Li, Eldon Y., Chien Hsiang Liao, and Hsiuju Rebecca Yen. 2013. "Co-Authorship Networks and Research Impact: A Social Capital Perspective." *Research Policy* 42 (9): 1515–30. https://doi.org/10.1016/j.respol.2013.06.012.

Lutter, Mark, and Martin Schröder. 2016. "Who Becomes a Tenured Professor, and Why? Panel Data Evidence from German Sociology, 1980–2013." *Research Policy* 45 (5): 999–1013. https://doi.org/10.1016/j.respol.2016.01.019.

Moody, James. 2004. "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review* 69 (2): 213–38. https://doi.org/10.1177/000312240406900204.

Ossenblok, T. L. B., T. C. E. Engels, and G. Sivertsen. 2012. "The Representation of the Social Sciences and Humanities in the Web of Science–a Comparison of Publication Patterns and Incentive Structures in Flanders and Norway (2005-9)." *Research Evaluation* 21 (4): 280–90. https://doi.org/10.1093/reseval/rvs019.

Strauss, David, and Michael Ikeda. 1990. "Pseudolikelihood Estimation for Social Networks." *Journal of the American Statistical Association* 85 (409): 204–12. https://doi.org/10.1080/01621459.1990.10475327.

Universität Bremen. 2022. "Welcome at the SOCIUM." https://www.socium.uni-bremen.de/home/en/?

van Rijnsoever, Frank J., and Laurens K. Hessels. 2011. "Factors Associated with Disciplinary and Interdisciplinary Research Collaboration." *Research Policy* 40 (3): 463–72. https://doi.org/10.1016/j.respol.2010.11.001.

Wallace, Matthew L., Vincent Larivière, and Yves Gingras. 2012. "A Small World of Citations? The Influence of Collaboration Networks on Citation Practices." Edited by Petter Holme. *PLoS ONE* 7 (3): e33339. https://doi.org/10.1371/journal.pone.0033339.

Wuchty, S., B. F. Jones, and B. Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–39. https://doi.org/10.1126/science.1136099.