This is a post-print version of the accepted article: 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*.

Polarization in the Social Sciences: Assortative Mixing in Social Science Collaboration Networks is Resilient to Interventions

Philip Leifeld

University of Glasgow, Adam Smith Building, 40 Bute Gardens, Glasgow, G12 8RT, UK

Abstract

Academic collaboration in the social sciences is characterized by a polarization between hermeneutic and nomological researchers. This polarization is expressed in different publication strategies. The present article analyzes the complete co-authorship networks in a social science discipline in two separate countries over five years using an exponential random graph model. It examines whether and how assortative mixing in publication strategies is present and leads to a polarization in scientific collaboration. In the empirical analysis, assortative mixing is found to play a role in shaping the topology of the network and significantly explains collaboration. Co-authorship edges are more prevalent within each of the groups, but this mixing pattern does not fully account for the extent of polarization. Instead, a thought experiment reveals that other components of the complex system dampen or amplify polarization in the data-generating process and that microscopic interventions targeting behavior change with regard to assortativity would be hindered by the resilience of the system. The resilience to interventions is quantified in a series of simulations on the effect of microscopic behavior on macroscopic polarization. The empirical study controls for geographic proximity, supervision, and topical similarity (using a vector space model), and the interplay of these factors is likely responsible for this resilience. The paper also predicts the co-authorship network in one country based on the model of collaborations in the other country.

Keywords: scientific collaboration, co-authorship network, polarization, assortative mixing, exponential random graph model, network intervention, social sciences

1. Collaboration and Polarization in the Social Sciences

To harness the innovative potential of academic research, we need to understand how the social aspects of scientific collaboration work. The structure of scientific collaboration is often analyzed by drawing on proxy measures such as co-authorship networks. A co-authorship network is comprised of researchers as vertices and joint publications as edges [1, 2, 3, 4, 5, 6, 7, 8]. Modeling the structure of co-authorship networks can be instrumental for designing better higher education and research institutions, mapping knowledge domains [9, 10, 11, 12], understanding how innovation comes about [13, 14], and understanding in how far the network structure has repercussions on individual researchers' performance [15, 16, 2]. To this end, previous research finds that co-authorship networks are usually small-world networks with small average path distances [1, 17], are composed of empirically measurable communities revolving around (sub-)disciplines [18, 19, 20, 21] and national science systems [22], and exhibit strong degree heterogeneity with few "star" researchers who have very active and diverse co-authorship profiles [23].

Yet, different scientific disciplines exhibit somewhat different co-authorship patterns [24]. While the natural sciences often follow a "lab model" with many participating researchers at different seniority levels listed as co-authors on publications, the modal number of authors on a paper in mathematics or in the social

Email address: philip.leifeld@glasgow.ac.uk (Philip Leifeld)

sciences is one [18]. Science, technology, engineering and math (STEM) fields almost exclusively value journal publications whereas monographs and edited volumes are relatively popular publication forms in the social sciences and especially the humanities. Existing research on co-authorship networks tends to focus on STEM fields because data on journal publications are easily available from electronic databases [25, 26, 27, 28]. Social science co-authorship network studies are therefore relatively rare (but see [23, 29, 30, 31, 32]), and those analyses that do exist are often biased in favor of journal articles because they, too, draw on journal publication databases. This practice yields a biased image of the scientific collaboration patterns of some disciplines like the social sciences and humanities, where journal publications are only one part of the picture [19, 20].

Few existing studies employ inferential network models to identify the generative mechanisms of scientific collaboration (notable exceptions: [33, 34, 35]). Those studies that do are likely to suffer from omitted variable bias because they focus on specific parts of the data-generating process while omitting other, potentially important sources of variation in edge formation, for example topic similarity in the research portfolios of authors, geographic proximity, social organization, and differential publication strategies. This study rectifies this situation by estimating a full-fledged inferential network model to the mechanics of scientific co-authorship, with extensive goodness-of-fit assessment in order to evaluate the macroscopic consequences of individual co-authorship behavior in comparison to the observed network.

In particular, restricting analysis of social-scientific collaboration patterns to journal publications, as is common practice in scientometrics, misses an important part of the data-generating process at the microscopic level: there exist different research traditions that allegedly tend to fence themselves off against external influences, and these groups of researchers with different research traditions pursue very different publication strategies. As Habermas [36] puts it,

> "*The nomological sciences, whose aim it is to formulate and verify hypotheses concerning the laws governing empirical regularities, have extended themselves far beyond the sphere of the theoretical natural sciences, into psychology and economics, sociology and political science. On the other hand, the historical-hermeneutic sciences, which appropriate and analyze meaningful cultural entities handed down by tradition, continue uninterrupted along the paths they have been following since the nineteenth century. There is no serious indication that their methods can be integrated into the model of the strict empirical sciences. [. . .] The analytic school dismisses the hermeneutic disciplines as prescientific, while the hermeneutic school considers the nomological sciences as characterized by a limited preunderstanding* [36, p. 1–2]."

By analyzing journal publications as available through online databases, existing network analyses of co-authorship networks in the social sciences tend to focus on the nomological camp almost exclusively [1, 18] and thereby entirely miss out on this mutual fencing off between the camps as an important part of the data-generating process (see also [37]).

In the present contribution, the focus is therefore on assortative mixing along the two research traditions in order to assess the extent of polarization and its resilience in the data-generating process. Polarization is a theoretical concept that describes the partitioning of vertices into clusters based on assortativity on a vertex variable, here publication strategies as a proxy for approaches to research. The stronger the assortativity, the higher the polarization of the network into two camps. This article answers three interrelated questions: do the two alleged camps really exist empirically, how entrenched are the two camps, and would an intervention be able to change the extent of polarization between them?

These research questions are answered by i) analyzing the complete political science co-authorship networks in two separate countries, including all types of publications such as book chapters in edited volumes; ii) estimating a state-of-the-art generative model of local interactions among researchers and evaluating its macroscopic consequences for the topology of the national co-authorship network at large; iii) including a full range of control factors that have been ignored in previous applications of statistical models to the mechanics of scientific collaboration, including geographic distance, supervisory relationships, and topic similarity; iv) predicting the complete co-authorship network in one country out of sample based on the model estimated with data in the other country, thereby cross-validating the model; v) examining the effect of assortative

mixing conditional on different levels of behavior of the two vertices within a researcher dyad; and vi) simulating new co-authorship networks with lower or higher levels of polarization between the two camps as thought experiments in order to evaluate the macroscopic effects of increased versus decreased polarizing behavior at the miscroscopic level. This serves to understand the extent of resilience of assortative mixing around the two camps.

2. Exponential random graph model

The co-authorship network is modeled using an exponential random graph model (ERGM) [38, 39, 40, 41, 42, 43]. The ERGM is a principled way of modeling the topology of a network based on covariates alongside endogenous dependencies. In contrast to growth and preferential attachment models, which have been found to explain the topology of scientific collaboration networks [44, 45, 46], the ERGM permits modelling the macroscopic properties of the system as a consequence of interacting microscopic properties that include local graph processes and exogenous vertex-related and dyadic factors, and to test the significance of these factors explicitly.

The probability density function of the exponential random graph model can be expressed as

$$\mathcal{P}(M, \boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(M)\}}{\sum_{M^* \in \mathcal{M}} \exp\{\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(M^*)\}}, \tag{1}$$

where $\boldsymbol{\Gamma}(M)$ can be arbitrary functions of the network M; this is where dependencies or covariates enter the model in the form of subgraph products. $\boldsymbol{\theta}$ is a vector of parameters to be estimated. Equation 1 reflects the probability of observing a specific topology M of the network over the networks $M^* \in \mathcal{M}$ one could have observed [38, 40]. $\boldsymbol{\Gamma}$ can also contain exogenous data such as functions of the original bipartite network N or covariates X. The model is estimated by Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC-MLE) [42].

The simplest case of a subgraph product entering the model through $\boldsymbol{\Gamma}(M)$ is the number of edges, the `edges` term,

$$\boldsymbol{\Gamma}_{\text{edges}}(M) = \sum_{i \neq k} M_{ik}, \tag{2}$$

where i and k are distinct vertices. Similarly, a dyadic covariate is defined as

$$\boldsymbol{\Gamma}_{\text{edgecov}}(M, X) = \sum_{i \neq k} X_{ik} M_{ik}, \tag{3}$$

where X_{ik} denotes the covariate value of dyad (i, k). Vertex covariates can be expressed as

$$\boldsymbol{\Gamma}_{\text{vertexcov}}(M, \vec{\mathbf{x}}) = \sum_{i \neq k} x_i M_{ik}, \tag{4}$$

where $\vec{\mathbf{x}}$ is a covariate vector of length m. Another subgraph product is a `vertexmatch` term, which captures how many (i, k) dyads have a match on vertex covariate $\vec{\mathbf{x}}$:

$$\boldsymbol{\Gamma}_{\text{vertexmatch}}(M, \vec{\mathbf{x}}) = \sum_{i \neq k} [x_i = x_k] M_{ik}, \tag{5}$$

where $[\ldots]$ denotes Iverson brackets, i. e., the term is 1 if the condition in brackets is true and 0 otherwise.

These exogenous covariate terms (Equations 3 to 5) serve to test substantive hypotheses on edge formation. Additional *endogenous* processes can be specified as subgraph products, for example triadic closure, cycles of arbitrary size, or k-stars [42, 43], in order to test relational hypotheses and/or rule out a violation of the i.i.d. assumption inherent in generalized linear models [38].

The "dependent variable" is the outcome network M or, if interpreted at the microscopic level, the binary state of each (i, k) dyad. The coefficients can be interpreted as log odds, conditional on the rest of the network, like in a logistic regression framework (the logit model is in fact a special case of the ERGM; see [39]).

	Germany		Switzerland	
	$w \geq 1$	$w \geq 2$	$w \geq 1$	$w \geq 2$
Number of vertices in M	1322	1322	156	156
Largest component	674	486	89	27
Number of edges in M	1329	1019	160	92
Average degree	2.01	1.54	2.05	1.18
Global clustering	0.39	0.44	0.26	0.31
Average local clustering coefficient	0.05	0.05	0.06	0.04
Modularity (Louvain [48])	0.86	0.89	0.75	0.86
Components larger than one	74	87	12	20
Average size of components	11.92	8.48	9.75	4.70
Average geodesic distance	7.77	8.10	4.79	2.58
Publications	22080		1904	
Journal articles	6518		732	
Book chapters	8289		546	
Monographs	1238		100	
Edited volumes	1516		56	
Institutions (u)	97		12	

Table 1: Summary statistics for the two datasets. $w \geq 1$ denotes the unweighted network; $w \geq 2$ denotes omission of edges with a weight of one.

3. Constructing two datasets

Two national case studies are analyzed. One of them is the complete co-authorship network among political scientists in Germany at the end of 2013 and reported retrospectively for the last five years [20]. The other one is the same kind of network for Switzerland [19]. The two countries have relatively balanced numbers of nomological and hermeneutic researchers (in contrast to, for example, the United States, where the nomological camp outnumbers the hermeneutic camp, or the United Kingdom, where the reverse pattern can be found). The two countries are also sufficiently similar for a direct comparison, with the two main differences being geographical size (such that geographical distance may play less of a prohibitive role for collaboration in Switzerland) and multilingualism.

Usually, co-authorship datasets are collected from electronic databases. However, these databases usually ignore non-peer-reviewed publications, which account for a substantial number of publications in the social sciences in Europe. To get a full picture of the discipline in the respective country, a manual three-step data collection strategy was pursued: first, a list of all $u = 97$ university departments and research institutes related to political science or public policy in Germany was created ($u = 12$ in Switzerland). Second, a table of all $m = 1{,}322$ researchers with a PhD listed on the homepages of these institutions was created ($m = 156$ in Switzerland). Third, citation details of all $n = 22,080$ publications of these researchers ($n = 1,904$ in Switzerland) in the last five years were collected from their institutional and private homepages and CVs. In a second round at the end of 2014, publications were added that had not been reported on the homepages or CVs immediately. More details on the data collection and quality can be found in [19, 20].

Of the $n = 22{,}080$ publications in Germany, 6,518 are journal articles, 8,289 are book chapters, 1,238 are monographs, 1,516 are edited volumes, and 4,519 are other kinds of publications. The proportions are similar for Switzerland (see [19]). These counts include duplicate publications that were reported by multiple researchers, but the network is constructed as binary, which renders duplicates inconsequential for the analysis [47].

An $m \times n$ bipartite network matrix N was created with 1 indicating the presence and 0 the absence of authorship of a document by author i. The co-authorship co-occurrence matrix to be modeled as the outcome network is the $m \times m$ row-based projection $M := NN^T$, which indicates which researcher vertices

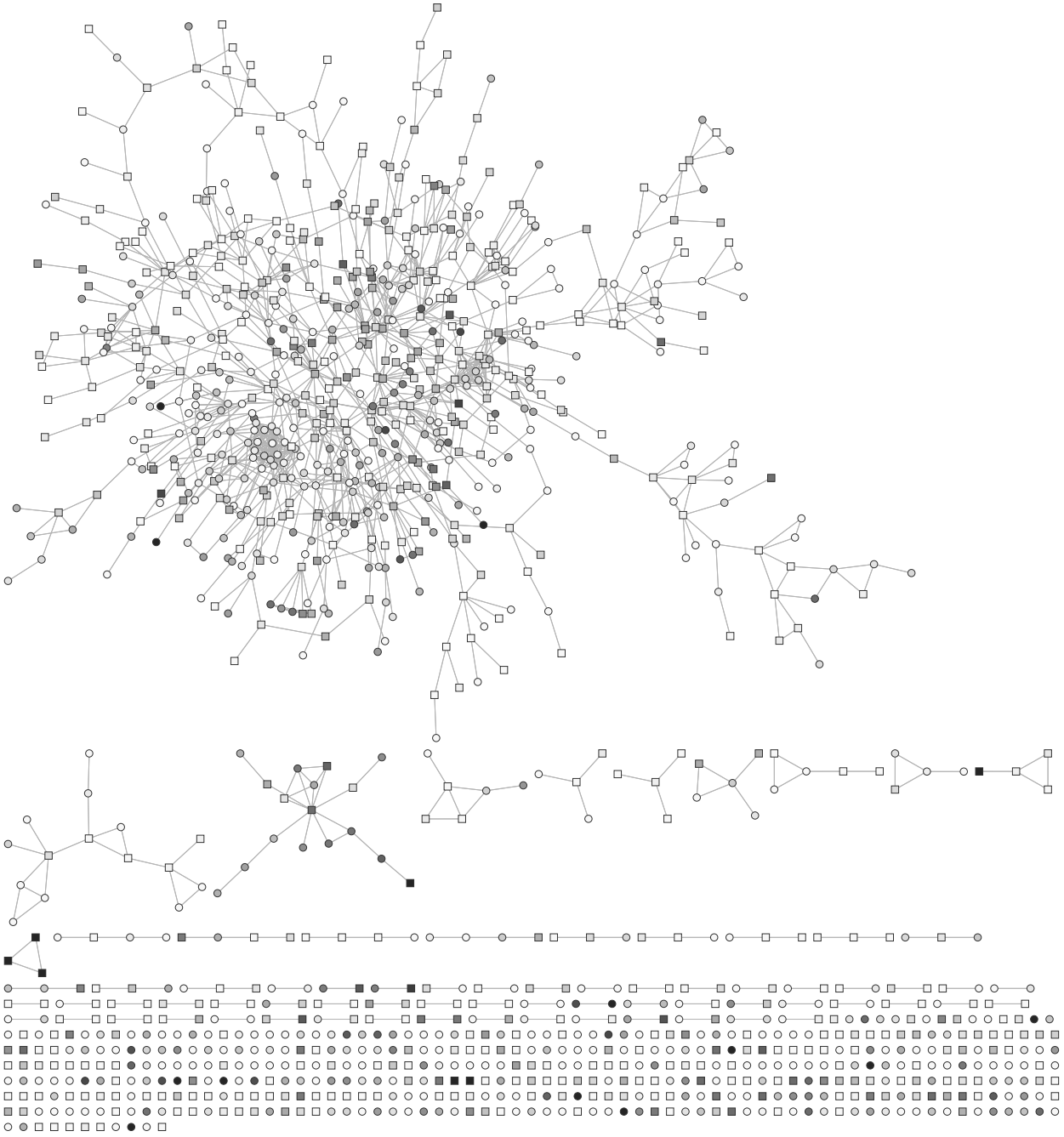

Figure 1: The German political science co-authorship network (1,322 vertices and 1,329 edges). Squares are professors and circles are postdocs or senior researchers. Color intensity from white to black indicates share of English articles (from low to high). The network exhibits moderate assortativity in terms of publication strategies.

i and k are connected by a co-authorship edge (see [49, 50] for an overview of network science concepts). A threshold value w is used to binarize M. w allows for the analysis of all collaborations (i. e., $w \geq 1$) or, as a robustness check, only intense collaboration (e. g., $w \geq 2$). These two threshold values are the two most frequent values in the distribution of edge weights. Figure 1 shows the co-authorship network in

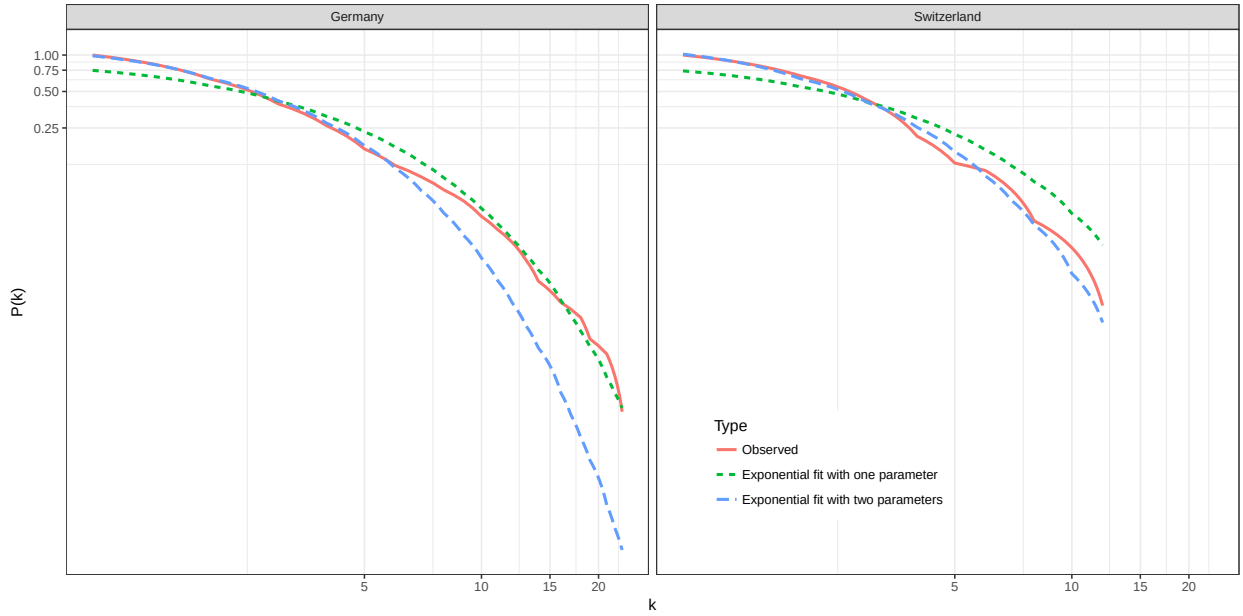

Figure 2: Log-log plot of the cumulative degree distribution of the largest component of the unweighted German and Swiss co-authorship networks, along with predicted values of fitted exponential models with one parameter ($f(k) = e^{-\gamma k} + \epsilon$ with $\gamma = 0.29$ for Germany and $\gamma = 0.30$ for Switzerland, in green) and two parameters ($f(k) = \alpha + e^{-\gamma k} + \epsilon$ with $\alpha = 0.41$ and $\gamma = 0.43$ for Germany and $\alpha = 0.48$ and $\gamma = 0.46$ for Switzerland, in blue). Degree on the x axis; cumulative relative frequency on the y axis.

Germany at $w \geq 1$ (i. e., using the unweighted graph). Figure 2 shows a log-log plot of the cumulative degree distributions of the giant components of the unweighted Swiss and German networks, along with predicted values of a fitted exponential model. Table 1 lists summary statistics for all networks.

4. Polarization and assortative mixing

Polarization between the nomological and the hermeneutic schools of thought is operationalized through their differing publication strategies. This assumes that nomological research is predominantly published in international academic journals and in English while hermeneutic research is more traditional and hence predominantly published in non-peer-reviewed outlets (e. g., books and book chapters) and in the local language (German in the German network and German, French, or Italian in the Swiss network).

Polarization is a macroscopic feature of a network. A polarized network contains a community structure, in this case with regard to an attribute of the vertices: the share of publications of a researcher in international journals and in English. A polarized network is composed of one faction with predominantly high values of this attribute and one faction with predominantly low values, and the density of connections within each faction is higher than between factions.

At the microscopic level, polarization can be operationalized by assortative mixing among vertices according to their attribute values [51, 41]. In an ERGM, assortative mixing can be modeled as a count of absolute differences on an attribute between all pairs of connected vertices. Since this is an interaction effect between sender and receiver, the main effects for the sender attribute and the receiver attribute must be included to achieve a straightforward interpretation.

More formally, the two hypotheses are tested as follows. Let $\vec{\mathbf{e}}$ be a vector with n elements where $e_j = 1$ if publication j is written in English and $e_j = 0$ otherwise. Let $\vec{\mathbf{a}}$ be a vector with n elements where $a_j = 1$ if publication j is a journal article and $a_j = 0$ otherwise. Then the share of English journal articles over

researcher i's publications amounts to

$$\mathrm{SEA}_i = \frac{\sum_{j=1}^{n} N_{ij} e_j a_j}{\sum_{j=1}^{n} N_{ij}}. \tag{6}$$

This quantity is included as a vertex covariate (Equation 4), and a positive coefficient is expected.

To compute the similarity between researchers i's and k's SEA values, one has to focus only on extra-dyadic publications because including joint publications in the similarity measure would mean introducing endogenous information from the dependent variable into the covariate. Based on this argument, English article similarity EAS_{ik} between researchers i and k is defined as

$$\mathrm{EAS}_{ik} = 1 - \left| \frac{\sum_{j=1}^{n} N_{ij}(1-(N_{ij}N_{kj}))e_j a_j}{\sum_{j=1}^{n} N_{ij}(1-(N_{ij}N_{kj}))} - \frac{\sum_{j=1}^{n} N_{kj}(1-(N_{ij}N_{kj}))e_j a_j}{\sum_{j=1}^{n} N_{kj}(1-(N_{ij}N_{kj}))} \right| \tag{7}$$

and included in the model as a dyadic covariate (Equation 3). This similarity index is defined in the range of $[0; 1]$; its mean is 0.79, and its median is 0.86, with a standard deviation of 0.22.

5. Control variables

5.1. Complementary skills and shared workload

How else can co-authorship behavior be explained? A few control variables are necessary to avoid omitted variables bias. The simplest explanation is that researchers have complementary skills or want to share the workload [52]. The consequence is that publications often have more than one author, and sometimes more than two (for an analysis of the number of authors, see [53]). In the co-authorship network, this leads to local configurations where edges between two researchers tend to be complemented by shared partners (leading to triadic closure or local clustering), but the number of shared partners per edge decays exponentially (meaning that one or two shared partners per connected dyad are much more frequent than, say, five or six common neighbors). Therefore, an endogenous network statistic called the geometrically weighted edge-wise shared partner distribution (GWESP) (see Equation 25 in [54]) is included in the model. The functional form captures the tendency that researchers who are co-authors have multiple other shared co-authors but that the likelihood of additional shared co-authors is exponentially decaying:

$$\mathbf{\Gamma}_{\mathrm{GWESP}}(M, \alpha) = \mathrm{e}^{\alpha} \sum_{h=1}^{m-2} \left\{1 - (1 - \mathrm{e}^{-\alpha})^h\right\} \mathrm{ESP}_h(M), \tag{8}$$

where $\mathrm{ESP}_h(M)$ is the number of edges that have exactly h shared partners and α determines the functional shape (0.3 for $w \geq 1$ and 0.2 for $w \geq 2$, as determined by model fit).

5.2. Seniority and productivity

Moreover, some researchers have more publications than other researchers and therefore have a greater chance of co-authoring with an arbitrary alter. Publication output is a function of three mechanisms: seniority, academic age, and different publication strategies: first, the more senior the position of a researcher, the more others seek to co-author with the researcher and the more "coordinative" authorships the researcher holds. Second, the longer ago the researcher obtained his or her PhD, the more publications and thus potential for co-authorship edges the researcher has accumulated. Third, there may be different publication strategies (peer-reviewed versus non-peer-reviewed), and it is possible that the number of peer-reviewed journal articles one can produce per time unit is lower than the number of non-peer-reviewed book chapters one can produce per time unit.

To control for these differential output frequencies, three model terms are included: first, the number of publications of a researcher (`Publication frequency`) is a direct control (as a vertex covariate as in Equation 4); second, a dummy variable (`Professor`) controls for the seniority level. Let $\vec{\mathbf{s}}$ be a vector of

seniority levels for all researchers, where $s_i = 1$ denotes a professor and $s_i = 0$ denotes a researcher who is not a professor but holds a PhD. $\vec{\mathbf{s}}$ is included in the model as a vertex covariate (Equation 4). Third, the number of co-authors researchers have across the network forms an exponentially decaying distribution, with few researchers having many co-authors and vice-versa [23]. This distribution is captured by an endogenous dependency term called the geometrically weighted degree distribution (`GWD`) (see Equation 14 in [54]),

$$\mathbf{\Gamma}_{\mathrm{GWD}}(M, \lambda) = \mathrm{e}^{\lambda} \sum_{h=1}^{m-1} \left\{1 - (1 - \mathrm{e}^{-\lambda})^h\right\} \mathrm{D}_h(M), \tag{9}$$

where $\mathrm{D}_h(M)$ is the number of vertices that have degree centrality h. λ is set to 0.4 for $w \geq 1$ and 0.6 for $w \geq 2$ on the basis of model fit. While $\mathbf{\Gamma}_{\mathrm{GWD}}(N, \lambda)$ effectively controls for the distribution of researchers' activity—e. g., with junior researchers having fewer co-authors than senior scholars—, $\mathbf{\Gamma}_{\mathrm{GWESP}}(M, \alpha)$ accounts for the fact that publications often have more than two authors, which leads to triadic closure. Controlling for these network dependencies is an important part of capturing the data-generating process.

5.3. Affiliation, chair, supervision, and geography

In addition to seniority, proximity mechanisms may play an important role. Collaboration likely takes place within teams, and it is especially likely that postdocs publish with their professor by whom they are paid and supervised. Even beyond team work, researchers from the same university or institute may be more likely to know each other, and mutual awareness may raise the probability of collaboration via acquaintance. These mechanisms are accommodated in the model as follows.

Let U be an $m \times u$ binary matrix with affiliations of researchers i or k to university or employer j. Let $\vec{\mathbf{c}}$ be a vector with m names of chairs or research groups of researchers i or k. Then shared university or employer affiliations between researchers are computed as the projection $UU^{\top}$ and introduced into the model as a dyadic covariate (Equation 3). Multiple shared affiliations are possible. Moreover, I construct a `vertexmatch` term (see Equation 5) using the $\vec{\mathbf{c}}$ vector for inclusion in the model, and I add an interaction term between seniority status, chair, and affiliation, which captures supervision of postdocs by professors within research teams at the same university:

$$\mathbf{\Gamma}_{supervision}(M, U, \vec{\mathbf{c}}, \vec{\mathbf{s}}) = \sum_{i \neq k} M_{ik} [\sum_{j=1}^{u} (U_{ij} U_{kj}) > 0] s_i (1 - s_k) [c_i = c_k] \tag{10}$$

Finally, while joint affiliations account for acquaintances and mutual awareness, collaboration is easier the smaller the geographic distance between two researchers, even across universities. Therefore a dyadic covariate with the geographic distance computed over the latitude and longitude of the institutes of the respective researchers is included in the model.

5.4. Topic similarity

In addition to these social factors, the actual contents of publications are presumably important. If researchers i and k work on similar topics, they are more likely to collaborate [55]. To include a topic similarity matrix between researchers as a covariate in the ERGM, I employ a vector space model [56].

Before proceeding, all publication titles are preprocessed by removing stop words and stemming the remaining words. Stop words are frequent words like "and" or "in". First, the language of each publication title is identified using the `jlangdetect` Java library.[1] Second, stop words are removed in seven languages that cover almost all titles (German, English, French, Italian, Spanish, Russian, and Dutch).[2] Third, the remaining words are stemmed in these seven languages using the Porter stemming algorithm[3] [57] in order to

[1] `https://github.com/melix/jlangdetect` (accessed 9 December 2017).

[2] Stop word lists from `snowball.tartarus.org/algorithms/dutch/stop.txt` (accessed 9 December 2017) for Dutch and from `http://members.unine.ch/jacques.savoy/clef/` (accessed 9 December 2017) for the remaining six languages are employed.

[3] `https://tartarus.org/martin/PorterStemmer/` (accessed 9 December 2017).

make words with different suffixes comparable. A vector space model is applied to the resulting preprocessed titles.

Let T be a weighted $p \times n$ matrix indicating how many times a word indexed in row l of the matrix shows up in the title of the publication indexed in column j. Then the term frequency

$$\text{tf}_l = \sum_{j=1}^{n} [T_{lj} > 0] \tag{11}$$

measures in how many distinct publication titles term l shows up at least once.

Some terms like "politic" are substantively unimportant because they appear in many titles and thus have very little discriminatory power while other terms like "abortion" show up in few titles and thus have a strong discriminatory power. To weight the terms in the similarity matrix according to their discriminatory power, I compute inverse document frequencies

$$\text{idf}_l = \log \frac{n}{\text{tf}_l}, \tag{12}$$

which place a higher value on rare words [58].

For each researcher dyad (i, k), I compute the extra-dyadic term frequencies of researcher i, which are defined as the number of times researcher i uses term l across all n titles but not in joint publications with researcher k:

$$\text{tf}_{ikl} = \sum_{j=1}^{n} T_{lj} \cdot N_{ij} \cdot (1 - (N_{ij} \cdot N_{kj})) \tag{13}$$

Excluding joint publications ensures that no endogenous properties of the network go into the computation of the topic similarities later. Based on these weighted term frequencies, one can compute idf-weighted term frequencies [59]

$$\text{tf-idf}_{ikl} = \log(1 + \text{idf}_l \cdot \text{tf}_{ikl}). \tag{14}$$

Finally, the extra-dyadic topic similarity between any two researchers i and k is defined as the cosine similarity of the extra-dyadic idf-tf vectors of i and k, which is the dot product of the two vectors over all terms divided by the product of the norms of the two vectors [60]:

$$\cos\theta = \frac{\sum_l^p \text{tf-idf}_{ikl} \cdot \text{tf-idf}_{kil}}{\sqrt{\sum_l^p (\text{tf-idf}_{ikl})^2} \cdot \sqrt{\sum_l^p (\text{tf-idf}_{kil})^2}} \tag{15}$$

This topic similarity matrix is included in the ERGM as a dyadic covariate (see Equation 3). It captures the division into subfields but is more fine-grained. The minimum of the similarity variable is 0, the maximum is 0.49 (median: 0.01; mean: 0.02; standard deviation: 0.02).

5.5. Gender homophily

Gender homophily is modeled using a `vertexmatch` term (Equation 5). If researchers i and k are of the same gender, their co-authorship probability may be increased.

6. Empirical results

Tables 2 and 3 show the results for two threshold values: all co-authorship edges ($w \geq 1$) and intense co-authorship relations ($w \geq 2$, requiring at least two shared publications for a co-authorship edge to emerge), for the German and Swiss co-authorship network, respectively. Both thresholds yield nearly identical results, which implies that the patterns identified are resilient to decisions made at the data collection stage. The tables also contain a minimal model with the most important model terms identified through a leave-one-out procedure (Figure 5).

	All edges	Intense collaboration	Minimal model
Endogenous model terms			
Edges	-11.35 (0.27)***	-12.26 (0.34)***	-10.33 (0.22)***
Edge-wise shared partners	1.90 (0.06)***	2.08 (0.07)***	2.17 (0.04)***
Degree distribution	0.40 (0.08)***	0.31 (0.09)***	
Exogenous covariates			
Publication frequency	0.00 (0.00)***	0.01 (0.00)***	
Professor	0.18 (0.04)***	0.17 (0.04)***	
Gender: male	0.05 (0.04)	-0.02 (0.05)	
Gender homophily	0.24 (0.07)***	0.30 (0.08)***	
Geographic distance	-0.12 (0.02)***	-0.13 (0.03)***	
Same affiliation	1.44 (0.07)***	1.45 (0.08)***	
Same chair or team	1.27 (0.12)***	1.29 (0.13)***	
Supervision	0.40 (0.16)*	0.38 (0.18)*	
Topic similarity	23.15 (0.63)***	23.60 (0.70)***	24.68 (0.48)***
Share of English articles	0.24 (0.10)*	0.34 (0.12)**	-0.07 (0.08)
English article similarity	3.03 (0.26)***	3.68 (0.32)***	2.66 (0.23)***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Exponential random graph models for the German collaboration network.

	All edges	Intense collaboration	Minimal model
Endogenous model terms			
Edges	-10.07 (0.62)***	-12.19 (1.05)***	-8.01 (0.41)***
Edge-wise shared partners	1.51 (0.16)***	1.45 (0.21)***	1.46 (0.12)***
Degree distribution	1.00 (0.28)***	1.17 (0.38)**	
Exogenous covariates			
Publication frequency	0.02 (0.00)***	0.04 (0.01)***	
Professor	0.11 (0.12)	0.18 (0.19)	
Gender: male	0.36 (0.18)*	0.44 (0.25)	
Gender homophily	0.02 (0.23)	-0.05 (0.31)	
Geographic distance	-0.24 (0.15)	-0.40 (0.23)	
Same affiliation	1.04 (0.23)***	0.95 (0.34)**	
Same chair or team	1.36 (0.31)***	1.71 (0.41)***	
Supervision	0.51 (0.41)	0.73 (0.49)	
Topic similarity	15.50 (2.03)***	10.78 (2.83)***	18.81 (1.61)***
Share of English articles	0.40 (0.18)*	0.64 (0.27)*	0.09 (0.15)
English article similarity	2.03 (0.50)***	2.80 (0.75)***	2.46 (0.45)***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Exponential random graph models for the Swiss collaboration network.

After controlling for `Edge-wise shared partners` (GWESP) (= local clustering) and the `Degree distribution` (GWD), the exogenous covariates can be interpreted substantively.

Going from 0 to 1 on the `Share of English articles` drives up the odds of collaboration by roughly 63 percent ($100 \cdot (\exp(0.49) - 1)$) in the German case and 51 percent in the Swiss case, which is a hint that international journal publications require co-authorship relations to overcome the additional burden of the review process. For authors with a high share of international journal articles, sharing the workload and trading complementary skills is imperative.

More importantly, there is also a strong assortative mixing effect with regard to `English article similarity`. The more similar two actors are in their publication strategy, the more likely they are co-authors. Researchers with international, peer-reviewed journal publications tend to work with each other, and researchers focusing on publications in German and/or on books and book chapters tend to work together. This is strong evidence of a bifurcation of the discipline into traditionalist and progressive—or hermeneutic and nomological—scholars.

To provide a microscopic interpretation [61] of this result, Figure 3 presents predicted probabilities for different values of the `English article similarity` effect in Model 1 in the German case, conditional on combinations of the `Share of English articles` of the sender and receiver vertex. From the coefficient alone, one cannot tell if the assortative mixing is driven by the hermeneutic camp (i. e., both vertices are similar because they have a low share of English articles) or by the nomological camp (i. e., both vertices are similar because they have a high share of English articles). The right panel of Figure 3 shows the slope of the effect within the hermeneutic camp (i. e., both researchers are in the lower half of the distribution), within the nomological camp (i. e., both researchers in the upper half), and for mixed dyads. In all of these cases, there is a discernible positive effect, which means that higher similarity leads to a greater probability of co-authorship. The left panel shows the distribution of probabilities for each of these partitions, and they are all significantly larger than zero, with the greatest effect size, but also greatest variance, in the nomological camp.

`Topic similarity` exhibits a large and significant coefficient and can be considered one of the building blocks of scientific collaboration. The more similar the topics two researchers work on (not counting their joint publications), the more likely they are co-authors.

Controlling for the number of publications per researcher (`Publication frequency`) and seniority (`Professor`), there is assortative mixing with regard to gender—but only in the German case.

As expected, `Geographic distance` (in units of 100 km) decreases the probability of co-authorship between i and k. Living 100 km closer together means being about 12 percent more likely to be co-authors in the German network, all else being equal. As expected, geographic proximity does not matter significantly in the smaller country Switzerland.

In addition to geographic proximity, `Same affiliation` matters in both countries: authors from the same university are almost three (two in the Swiss case) times as likely to be co-authors as researchers from different institutes, controlling for common team membership and supervision. Being members of the `Same chair or team` increases the odds of collaboration by roughly 263 (286, respectively) percent, and—controlling for these relations—being in a supervision relation by 42 (67) percent.

Except for gender homophily and geographic proximity, the results lead to the same substantive conclusions in Germany and Switzerland.

7. Endogenous and predictive fit

Figure 4 shows indicators of the endogenous goodness of fit for $w \geq 1$ in Germany (with similar results for $w \geq 2$ and for Switzerland, which are not reported here). The gray boxplots are distributions of auxiliary network statistics like edge-wise shared partners (i. e., the distribution of the local clustering coefficient) or geodesic distances of artificial networks simulated from the estimated model. The median lines should be closely aligned with the red line, which represents the same statistics in the observed network. This is clearly the case; simulated co-authorship networks closely resemble the actual collaboration network on several important endogenous characteristics.

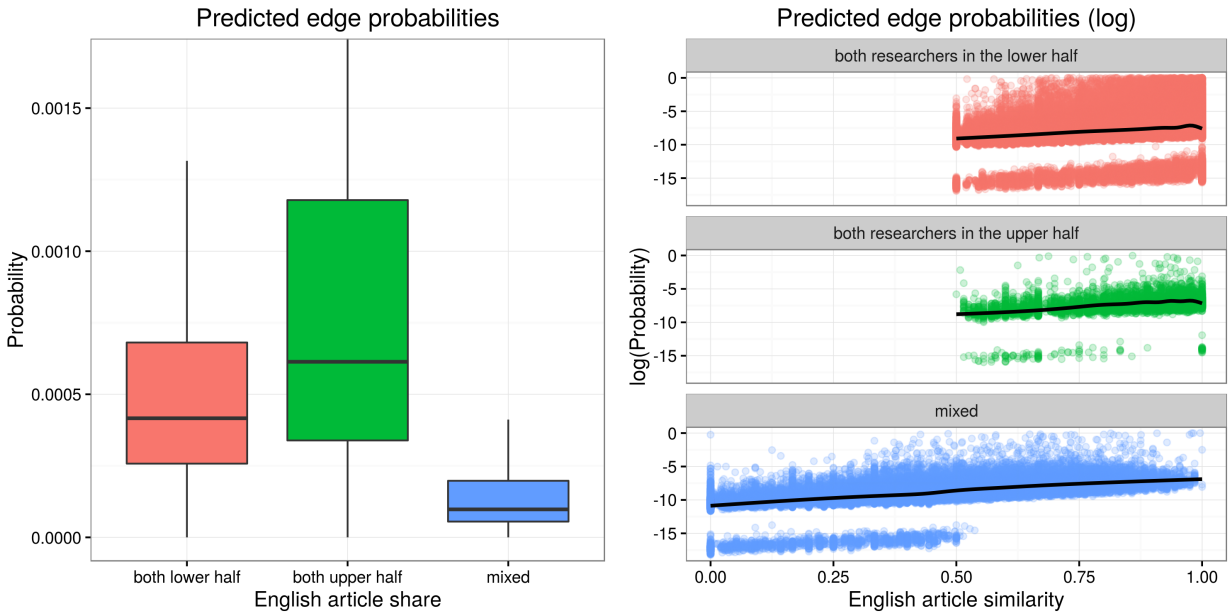

Figure 3: Microscopic interpretation of English article similarity. Left panel: Predicted probabilities if both authors are in the lower half, upper half, or in different halves of the `English article share` distribution. Probability of co-authorship is lowest for mixed dyads in which one author is on the nomological and the other author on the idiographic spectrum. Right panel: Distribution of edge probabilities (log) for different `English article similarity` values within each camp. Both within each group and across groups, there is a positive association between `English article similarity` and edge probability; the effect applies to both camps and does not focus on one end of the distribution.

Yet, this only measures how well the structural network properties are captured by the assumed data-generating process, not necessarily the actual location of the edges if the vertex labels matter. Figure 5 therefore assesses the classification performance of the first model, a complementary way of assessing the goodness of fit that considers the extent to which actual edges in the observed network are successfully predicted by the simulations. The precision–recall (PR) curves show how well the model performs overall with regard to recovery of the topology of the network [62].

The light red curve shows the model fit of a random graph with the same density. It serves as a null model against which the co-authorship model can be compared. The dark red and dark blue curves show the within-sample model fit of the $w \geq 1$ (i. e., unweighted) and $w \geq 2$ model, respectively. Both models work similarly well and perform much better than the null model. The model with intense edges ($w \geq 2$) performs slightly better than the full model ($w \geq 1$) based on within-sample edge prediction. This is only true in the German case; the opposite is true in the Swiss case.

The gray curves are the result of a leave-one-out algorithm. For each gray curve, the full model is recomputed, but one model term is left out at a time, and then PR curves are drawn. This procedure serves to assess the relative importance of each model term. The integral of the red curve minus the integral of the gray curve indicates the relative importance of the respective model term. The model terms with the greatest relative importance in and of themselves are `Edge-wise shared partners` (0.143), `Edges` (0.124), and `Topic similarity` (0.065). The general conclusion is that the model fit is the result of the complex interplay between multiple contributing factors and that each factor alone has limited explanatory power. The minimal specification reported in Tables 2 and 3 contain only these three important model terms and the two variables of primary substantive interest.

The green curve shows how well the model performs out of sample by predicting the Swiss co-authorship network as a test case based on the German training model with the Swiss covariates. The prediction is based on 100 simulations using a Metropolis–Hastings sampler [63, 64, 65] as implemented in the software `statnet`

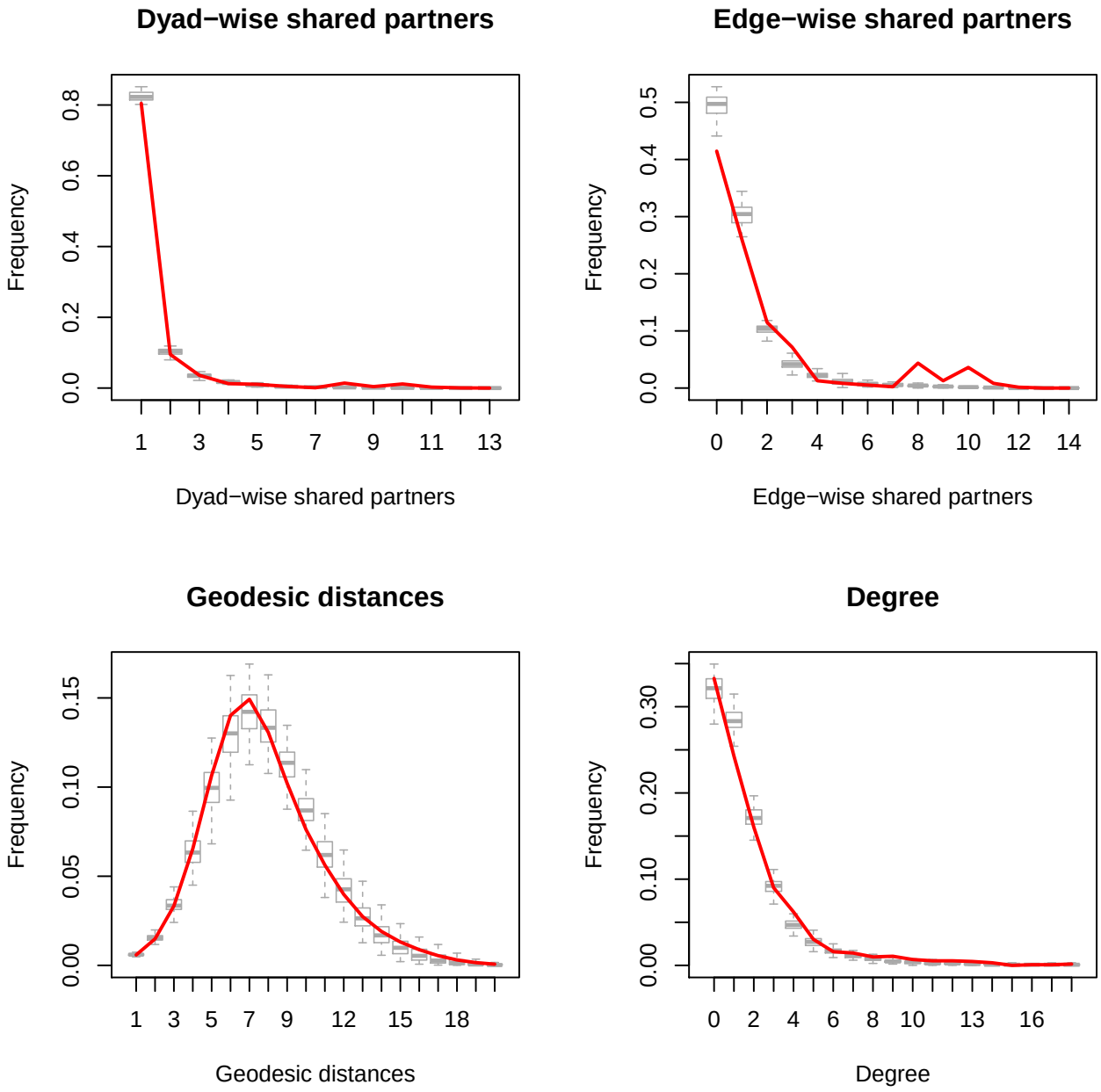

Figure 4: Endogenous goodness-of-fit assessment for $w \geq 1$: Comparison of several auxiliary network statistics between observed network (red lines) and 100 simulated networks (gray boxplots). Endogenous model fit is good if the red line and the median lines of the boxplots are nearly identical. Zero dyad-wise shared partners and infinite geodesic distances (= path distances) were omitted for reasons of legibility, but fitted well.

[66, 67]. Out-of-sample model fit based on the German model and Swiss data is almost as good as within-sample fit using the Swiss model, indicating that the model presented here has substantial explanatory and predictive value.

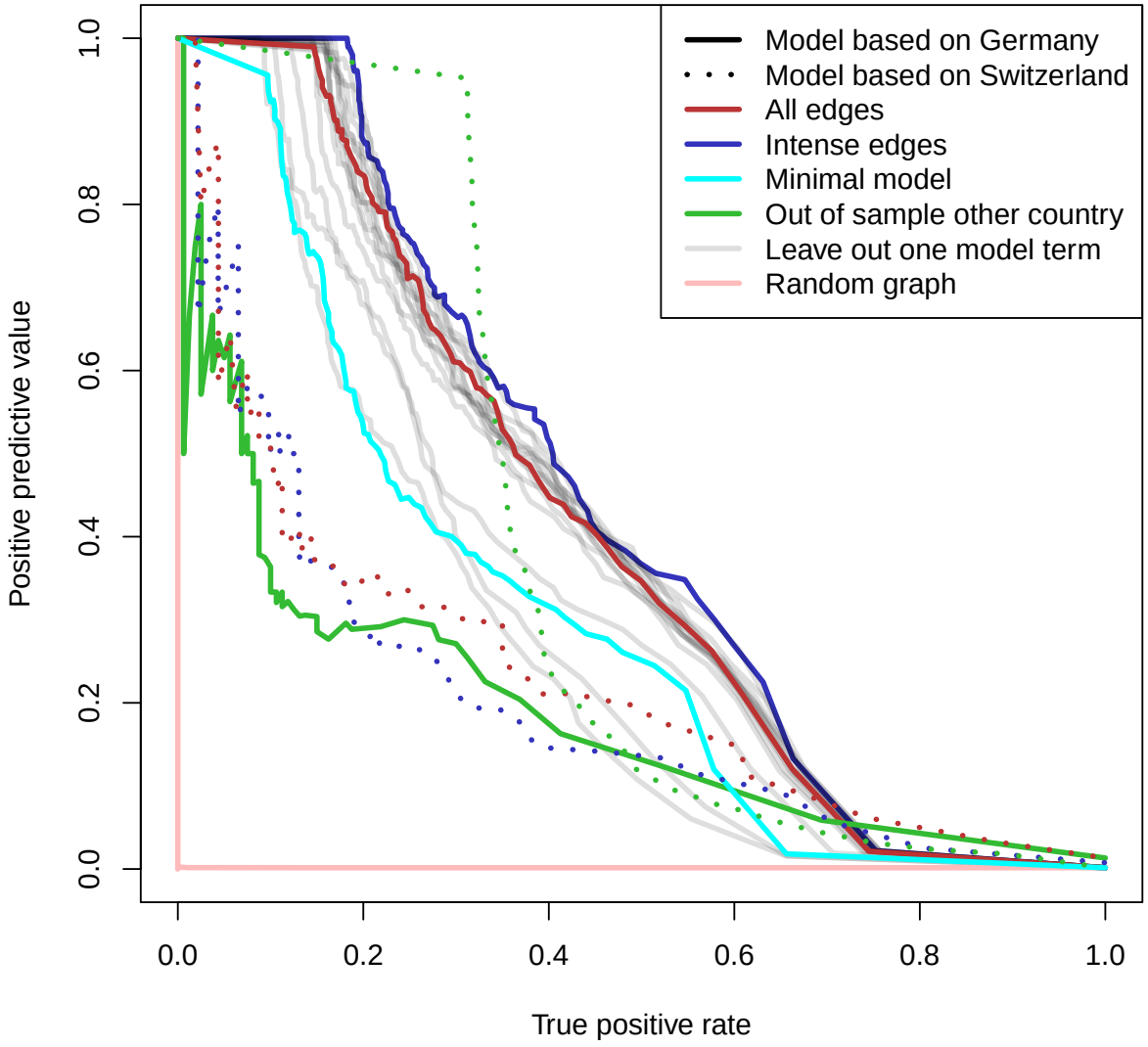

Figure 5: Prediction performance using precision–recall curves. Predictions based on the data from Germany are visualized as solid lines, predictions based on the data from Switzerland as dotted lines. Larger area under the curve indicates better predictive model fit (out-of-sample for the green line and within-sample for the other lines). Out-of-sample prediction of the Swiss network using coefficients estimated using data from Germany (solid green line) leads to a predictive fit that is comparable to the within-sample fit for the Swiss model; this increases confidence in the model.

8. Thought experiments on resilience to interventions

How consequential is the extent of assortative mixing by publication strategy for the overall polarization of the network, and how resilient or malleable is this polarization? Thought experiments on how different levels of individual assortative mixing would affect the overall mixing pattern in the population can provide an indication of the possible effects of any intervention or behavior change. If it *were* possible to change the

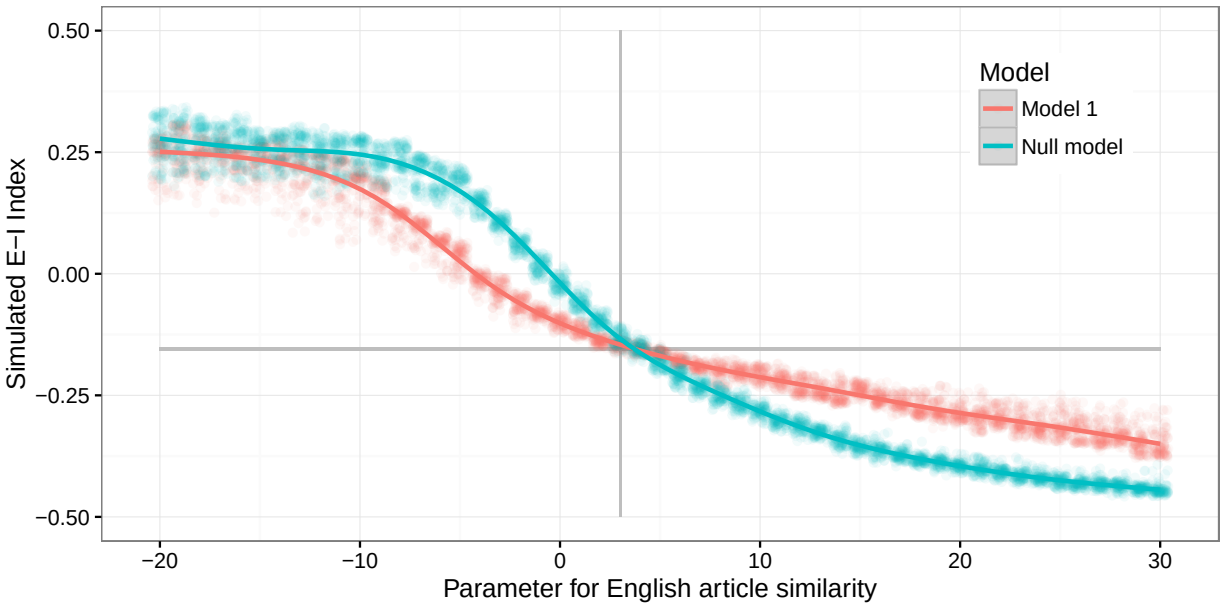

Figure 6: Simulation of E–I values for different parameters and models. MCMC-based simulations of networks with the same size and density as the observed network, varying only the coefficient for `English article similarity` (EAS) in the range of $[-20, +30]$. To offset polarization at the macroscopic level, an intervention would be required that changes the behavior of the individuals from EAS $= +3.03$ to EAS < -3.

average assortative mixing behavior of each individual represented in the network, how would this change the extent to which the overall network is fragmented?

While the network in Figure 1 exhibits a tendency for assortative mixing (see E-I index below), this pattern is not strong enough to partition the network into homogenous regions. Instead, small like-minded cohesive subgroups are intermingled with other like-minded subgroups of the opposite type, and there are frequent connections between them.

To evaluate assortative mixing more systematically, a Metropolis–Hasting sampler [63, 64, 65, 66, 67] was used to simulate networks of the same size and density as the observed network, based on the data-generating process and coefficients estimated in the first model, and the coefficient for `English article similarity` was systematically varied to assess its effects.

In this thought experiment, the `Share of English articles` variable was dichotomized at the median to produce two camps: those with above-median shares of English articles and those with below-median shares. The homogeneity of the two factions can be more concisely expressed by the E-I index [68], which measures the extent to which edges fall between, rather than within, two pre-defined groups:

$$\text{E-I index} = \frac{\text{EL} - \text{IL}}{\text{EL} + \text{IL}}, \tag{16}$$

where EL is the number of links external to the group and IL is the number of links internal to the group. The index is defined between $[-1, +1]$, with negative values indicating more homogenous factions and positive values indicating more heterogeneity.

Figure 6 shows the E–I index for additional simulations in the value range of $[-20, +30]$ at intervals of 1.0 for the `English article similarity` parameter (in red), with 100 simulated networks and resulting E–I values per parameter step. The original coefficient from Model 1 is highlighted with gray lines. The coefficient of 3.03 corresponds to an E–I value of -0.1546 and indicates a relative polarization, in line with the interpretation of the ERGM. With an increasing `English article similarity` parameter in the

ERGM-based simulations, the lower the E–I index gets and the stronger the division between the camps gets. With a decreasing parameter, the lower the polarization becomes.

However, in order to achieve zero polarization (i. e., an E–I value of 0 on the y axis), a negative parameter for `English article similarity` is necessary. This means that other interactions in the system account for part of the polarization, rather than assortative mixing alone. Negative assortative mixing with a parameter of < -3 can counterbalance these other polarizing forces.

Moreover, a comparison with a null model reveals the resilience of the system in terms of its polarization. The blue curve and points represent simulated E–I values based on an ERGM only with model terms for `Edges`, `Share of English articles` and `English article similarity` that was fitted to the same empirical data. The comparison between the two simulation conditions shows how the null model would much more quickly exhibit stronger polarization patterns with increasing parameter values of `English article similarity`. The point where the two curves are crossing each other is the estimated parameter in the null model. The integral of the red curve minus the integral of the blue curve equals the resilience of the complex system to interventions in assortative mixing patterns in the network. Left of the crossing point, much stronger interventions towards negative assortative mixing would be needed in the real world to offset the propensity of the system to exhibit polarization than in a random network of the same size and with the same degree of polarization. This can be considered a negative externality of the complex system. Right of the crossing point of the two curves, stronger assortative mixing leads to less polarization relative to the null model. In other words, the complex system is relatively resilient to additional marginal assortativity; polarization occurs at lower rates than in a network that is not governed by the remaining components, such as geographical proximity, topical similarity etc. This can be considered a positive externality of the complex system as the network prevents extreme polarization by design, even in situations where behavioral change is induced.

9. Conclusion

In this paper, I have presented an inferential network model of scientific collaboration in a social science discipline in two European countries. The focus was on assortative mixing in publication strategies and its possible consequences in terms of polarization.

The findings indicate that the alleged polarization between hermeneutic and nomological observed by philosophers of science can be identified empirically and traced back to assortative mixing among researchers along their differential publication strategies. However, the extent of polarization is much too small to cause a disconnect between the two alleged camps.

Macroscopic polarization is furthermore related to other microscopic factors than assortative mixing among researchers. A thought experiment revealed how negative assortative mixing would be needed to counterbalance these other factors, and how increased assortative mixing at the inter-individual level would marginally lead to less polarization than in a comparable network without those additional factors in the data-generating process.

These other microscopic factors that are relevant for polarization between hermeneutic and nomological researchers may be hidden in the interplay between topic similarity, geographic proximity, seniority, and supervision. It seems plausible that topical similarity, for example, is intrinsically linked to these camps. For example, a stronger role of topical similarity could lead to more isolated camps if the topical coherence only increased within these factions. Similarly, there may be geographical hotspots that are associated with each camp, or more senior academics may be more prone to be associated with the hermeneutic paradigm. There is statistical evidence that all of these microscopic factors together shape the topology of the network, but additionally at least some of them must dampen the effect of assortative mixing in publication strategies in producing or alleviating aggregate polarization. In other words, these features of the complex system make the system resilient to changes in assortativity behavior.

This prompts the question of transferability to other contexts. Since the findings robustly explain two entirely separate country networks in Europe in political science, it seems plausible that the substantive findings may extend to other social sciences in the same countries or political science in other countries.

Transferability may be hampered in countries with much more skewed distributions of relative numbers of nomological versus hermeneutic researchers, such as the United States and possibly the United Kingdom. It is unclear how and whether findings from this study hold in the STEM context.

On a more abstract level, this study shows how polarization in a complex social system can be caused by other factors than one would intuitively conceive. For example, in the study of legislative polarization or opinion polarization in a society [69, 70, 71, 72], macroscopic polarization of the network may be partly due to ideological differences, but may be partly also conditioned by factors such as geographically located incentives ("pork barrel politics") or additional, multiplex network relations [73]. This suggests that polarization in social or political settings may be much harder to understand or tackle than conventional wisdom would suggest. More research into intervention strategies in complex social systems is therefore warranted.

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

[2] Q. Shi, B. Xu, X. Xu, Y. Xiao, W. Wang, H. Wang, Diversity of social ties in scientific collaboration networks, Physica A: Statistical Mechanics and its Applications 390 (23) (2011) 4627–4635.

[3] J. A. Almendral, J. G. Oliveira, L. López, J. Mendes, M. A. Sanjuán, The network of scientific collaborations within the European Framework Programme, Physica A: Statistical Mechanics and its Applications 384 (2) (2007) 675–683.

[4] R. Lara-Cabrera, C. Cotta, A. J. Fernández-Leiva, An analysis of the structure and evolution of the scientific collaboration network of computer intelligence in games, Physica A: Statistical Mechanics and its Applications 395 (2014) 523–536.

[5] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications 311 (3) (2002) 590–614.

[6] A. Cardillo, S. Scellato, V. Latora, A topological analysis of scientific coauthorship networks, Physica A: Statistical Mechanics and its Applications 372 (2) (2006) 333–339.

[7] A. A. Roohi, A. H. Shirazi, A. Kargaran, G. R. Jafari, Local model of a scientific collaboration in physics network compared with the global model, Physica A: Statistical Mechanics and its Applications 389 (23) (2010) 5530–5537.

[8] Y. Li, C. You, What is the difference of research collaboration network under different projections: Topological measurement and analysis, Physica A: Statistical Mechanics and its Applications 392 (15) (2013) 3248–3259.

[9] L. M. A. Bettencourt, J. Kaur, Evolution and structure of sustainability science, Proceedings of the National Academy of Sciences 108 (49) (2011) 19540–19545.

[10] K. W. Boyack, Mapping knowledge domains: Characterizing PNAS, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5192–5199.

[11] D. A. Vilhena, J. G. Foster, M. Rosvall, J. D. West, J. Evans, C. T. Bergstrom, Finding cultural holes: How structure and culture diverge in networks of scholarly communication, Sociological Science 1 (2014) 221–239.

[12] K. K. Mane, K. Börner, Mapping topics and topic bursts in PNAS, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5287–5290.

[13] A. G. Vasconcellos, C. M. Morel, Enabling policy planning and innovation management through patent information and co-authorship network analyses: A study of tuberculosis in Brazil, PLOS ONE 7 (10) (2012) e45569. doi:10.1371/journal.pone.0045569.

[14] A. Pluchino, S. Boccaletti, V. Latora, A. Rapisarda, Opinion dynamics and synchronization in a network of scientific collaborations, Physica A: Statistical Mechanics and its Applications 372 (2) (2006) 316–325.

[15] S. Uddin, L. Hossain, K. Rasmussen, Network effects on scientific collaborations, PLOS ONE 8 (2) (2013) e57546. doi:10.1371/journal.pone.0057546.

[16] M. Prosperi, I. Buchan, I. Fanti, S. Meloni, P. Palladino, V. I. Torvik, Kin of coauthorship in five decades of health science literature, Proceedings of the National Academy of Sciences 113 (32) (2016) 8957–8962.

[17] M. Perc, Growth and structure of Slovenia's scientific collaboration network, Journal of Informetrics 4 (4) (2010) 475–482.

[18] M. E. J. Newman, Coauthorship networks and patterns of scientific collaboration, Proceedings of the National Academy of Sciences 101 (suppl. 1) (2004) 5200–5205.

[19] P. Leifeld, K. Ingold, Co-authorship networks in Swiss political research, Swiss Political Science Review 22 (2) (2016) 264–287.

[20] P. Leifeld, S. Wankmüller, V. T. Z. Berger, K. Ingold, C. Steiner, Collaboration patterns in the German political science co-authorship network, PLOS ONE 12 (4) (2017) e0174671.
URL http://dx.doi.org/10.1371/journal.pone.0174671

[21] B. Lužar, Z. Levnajić, J. Povh, M. Perc, Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network, PLOS ONE 9 (4) (2014) e94429. doi:10.1371/journal.pone.0094429.

[22] T. Vanni, M. Mesa-Frias, R. Sanchez-Garcia, R. Roesler, G. Schwartsmann, M. Z. Goldani, A. M. Foss, International scientific collaboration in HIV and HPV: A network analysis, PLOS ONE 9 (3) (2014) e93376.

[23] J. Moody, The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999, American Sociological Review 69 (2) (2004) 213–238.

[24] K. Jaffe, Social and natural sciences differ in their research strategies, adapted to work for different knowledge landscapes, PLOS ONE 9 (11) (2014) e113901. doi:10.1371/journal.pone.0113901.

[25] American Physical Society, APS data sets for research, Available online at https://journals.aps.org/datasets (accessed 25 March 2018) (2018).

[26] A. Çavuşoğlu, İ. Türker, Patterns of collaboration in four scientific disciplines of the Turkish collaboration network,

Physica A: Statistical Mechanics and its Applications 413 (2014) 220–229.
[27] A. Çavuşoğlu, İ. Türker, Scientific collaboration network of Turkey, Chaos, Solitons & Fractals 57 (2013) 9–18.
[28] J. C. Phillips, Phase transitions in the Web of Science, Physica A: Statistical Mechanics and its Applications 428 (2015) 173–177.
[29] F. J. Acedo, C. Barroso, C. Casanueva, J. L. Galán, Co-authorship in management and organizational studies: An empirical and network analysis, Journal of Management Studies 43 (5) (2006) 957–983.
[30] C. Gossart, M. Özman, Co-authorship networks in social sciences: The case of Turkey, Scientometrics 78 (2) (2009) 323–345.
[31] M.-G. Hâncean, M. Perc, L. Vlăsceanu, Fragmented Romanian sociology: Growth and structure of the collaboration network, PLOS ONE 9 (11) (2014) e113271. doi:10.1371/journal.pone.0113271.
[32] M.-G. Hâncean, M. Perc, Homophily in coauthorship networks of East European sociologists, Scientific Reports 6 (2016) 36152.
[33] L. Kronegger, F. Mali, A. Ferligoj, P. Doreian, Collaboration structures in Slovenian scientific communities, Scientometrics 90 (2) (2011) 631–647.
[34] X. Liang, The changing impact of geographic distance: A preliminary analysis on the co-author networks in scientometrics (1983–2013), in: 48th Hawaii International Conference on System Sciences (HICSS), IEEE, 2015, pp. 722–731.
[35] S. Zaccarin, G. Rivellini, Modelling network data: An introduction to exponential random graph models, in: F. Palumbo, C. N. Lauro, M. J. Greenacre (Eds.), Data Analysis and Classification, Springer, Berlin/Heidelberg, 2010, pp. 297–305.
[36] J. Habermas, On the Logic of the Social Sciences, John Wiley & Sons, 2015.
[37] M. Williams, L. Sloan, C. Brookfield, A tale of two sociologies: Analytic versus critique in UK sociology, Sociological Research Online (2017) 1360780417734146doi:10.1177/1360780417734146.
[38] S. J. Cranmer, P. Leifeld, S. D. McClurg, M. Rolfe, Navigating the range of statistical tools for inferential network analysis, American Journal of Political Science 61 (1) (2017) 237–251.
[39] B. A. Desmarais, S. J. Cranmer, Statistical mechanics of networks: Estimation and uncertainty, Physica A: Statistical Mechanics and its Applications 391 (4) (2012) 1865–1876.
[40] J. Park, M. E. J. Newman, Statistical mechanics of networks, Physical Review E 70 (6) (2004) 066117.
[41] A. L. Traud, P. J. Mucha, M. A. Porter, Social structure of Facebook networks, Physica A: Statistical Mechanics and its Applications 391 (16) (2012) 4165–4180.
[42] G. Robins, P. Pattison, Y. Kalish, D. Lusher, An introduction to exponential random graph (p*) models for social networks, Social Networks 29 (2) (2007) 173–191.
[43] S. Wasserman, P. Pattison, Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*, Psychometrika 61 (3) (1996) 401–425.
[44] H. Jeong, Z. Néda, A.-L. Barabási, Measuring preferential attachment in evolving networks, EPL (Europhysics Letters) 61 (4) (2003) 567. doi:10.1209/epl/i2003-00166-9.
[45] İ. Türker, A. Çavuşoğlu, Detailing the co-authorship networks in degree coupling, edge weight and academic age perspective, Chaos, Solitons & Fractals 91 (2016) 386–392.
[46] M. Perc, The Matthew effect in empirical data, Journal of The Royal Society Interface 11 (98) (2014) 20140378.
[47] M. Tomassini, L. Luthi, Empirical analysis of the evolution of a scientific collaboration network, Physica A: Statistical Mechanics and its Applications 385 (2) (2007) 750–764.
[48] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10) (2008) P10008.
[49] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.
[50] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports 424 (4) (2006) 175–308.
[51] M. E. J. Newman, Mixing patterns in networks, Physical Review E 67 (2) (2003) 026126.
[52] L. Wardil, C. Hauert, Cooperation and coauthorship in scientific publishing, Physical Review E 91 (1) (2015) 012825.
[53] S. Milojević, Principles of scientific research team formation and evolution, Proceedings of the National Academy of Sciences 111 (11) (2014) 3984–3989.
[54] D. R. Hunter, Curved exponential family models for social networks, Social Networks 29 (2) (2007) 216–230.
[55] K. Börner, J. T. Maru, R. L. Goldstone, The simultaneous evolution of author and paper networks, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5266–5273.
[56] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (11) (1975) 613–620.
[57] M. F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.
[58] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation 28 (1) (1972) 11–21.
[59] S. Robertson, Understanding inverse document frequency: On theoretical arguments for IDF, Journal of Documentation 60 (5) (2004) 503–520.
[60] S. K. M. Wong, W. Ziarko, P. C. N. Wong, Generalized vector spaces model in information retrieval, in: Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1985, pp. 18–25.
[61] B. A. Desmarais, S. J. Cranmer, Micro-level interpretation of exponential random graph models with application to estuary networks, Policy Studies Journal 40 (3) (2012) 402–434.
[62] J. Davis, M. Goadrich, The relationship between precision–recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240.

[63] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1) (1970) 97–109.
[64] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, The Journal of Chemical Physics 21 (6) (1953) 1087–1092.
[65] S. Chib, E. Greenberg, Understanding the Metropolis-Hastings algorithm, The American Statistician 49 (4) (1995) 327–335.
[66] D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, M. Morris, ergm: A package to fit, simulate and diagnose exponential-family models for networks, Journal of Statistical Software 24 (3) (2008) 1–29.
[67] M. Morris, M. S. Handcock, D. R. Hunter, Specification of exponential-family random graph models: Terms and computational aspects, Journal of Statistical Software 24 (4) (2008) 1–24.
[68] D. Krackhardt, R. N. Stern, Informal networks and organizational crises: An experimental simulation, Social Psychology Quarterly 51 (2) (1988) 123–140.
[69] J. H. Kirkland, Ideological heterogeneity and legislative polarization in the United States, Political Research Quarterly 67 (3) (2014) 533–546.
[70] D. R. Fisher, J. Waggle, P. Leifeld, Where does political polarization come from? Locating polarization within the US climate change debate, American Behavioral Scientist 57 (1) (2013) 70–92.
[71] P. Leifeld, Polarization of coalitions in an agent-based model of political discourse, Computational Social Networks 1 (1) (2014) 7.
[72] S. Galam, Public debates driven by incomplete scientific data: The cases of evolution theory, global warming and H1N1 pandemic influenza, Physica A: Statistical Mechanics and its Applications 389 (17) (2010) 3619–3631.
[73] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, The structure and dynamics of multilayer networks, Physics Reports 544 (1) (2014) 1–122.