Multilayer network approach to modeling authorship influence on citation dynamics in physics journals

Vahan Nanumyan ,[*] Christoph Gote ,[†] and Frank Schweitzer [‡]
Chair of Systems Design, ETH Zurich, 8092 Zurich, Switzerland

(Received 28 February 2020; accepted 6 August 2020; published 8 September 2020)

We provide a general framework to model the growth of networks consisting of different coupled layers. Our aim is to estimate the impact of one such layer on the dynamics of the others. As an application, we study a scientometric network, where one layer consists of publications as nodes and citations as links, whereas the second layer represents the authors. This allows us to address the question of how characteristics of authors, such as their number of publications or number of previous coauthors, impacts the citation dynamics of a new publication. To test different hypotheses about this impact, our model combines citation constituents and social constituents in different ways. We then evaluate their performance in reproducing the citation dynamics in nine different physics journals. For this, we develop a general method for statistical parameter estimation and model selection that is applicable to growing multilayer networks. It takes both the parameter errors and the model complexity into account and is computationally efficient and scalable to large networks.

DOI: 10.1103/PhysRevE.102.032303

I. INTRODUCTION

Citation networks and their growth long have attracted the interest of the complex networks research community [1–3]. A citation network is a directed network in which nodes represent scientific publications and edges between nodes citations between publications. Because new publications can only cite existing ones, the *temporal ordering* in which nodes and links are added to the network plays a crucial role, and a time-aggregated network representation omits this information.

One focus of current research was to test whether established growth mechanisms, such as preferential attachment, are suitable to describe the evolution of citation networks [4–6]. Preferential attachment assumes the growth rate, Δk, is proportional to the in-degree of a node, k, i.e., the number of citations a paper has attracted so far. This follows early investigations from the 1960s showing that publications with a large number of citations tend to attract even more citations [7,8].

Preferential attachment prescribes a certain average dynamics to every node, i.e., it does not account for any kind of *heterogeneity* other than the degree. In practice, however, some nodes, once added to the network, accumulate many edges very quickly, while others do not. This observation has led to the notion of node *fitness*: The fitter a node, the more edges it attracts [9–11]. Conceptually, this considers a heterogeneity among nodes that goes beyond network properties. In citation networks, the fitness is usually attributed to the content of the paper.

[*]vnanumyan@ethz.ch
[†]cgote@ethz.ch
[‡]fschweitzer@ethz.ch

Preferential attachment further implies that that *older* nodes attract more edges, which is in disagreement with empirical observations of citation dynamics. Hence, the *ageing* of nodes was introduced as another node property. With respect to publications, ageing reflects obsolescence (novelty decay, relevance decay) [12,13] or attention decay [2,14].

When verifying the preferential attachment hypothesis, most works either looked at the scale-free *degree distribution* as an aggregated property of the network or analyzed the relation between in-degree, k, and growth rate, Δk. These approaches suffer from certain drawbacks. For instance, the mentioned distributions and relations can result from many dynamic mechanisms, e.g., $\Delta k \sim k$ can be found from a preferential attachment or a fitness model [6]. In addition, results on statistical modeling are not conclusive without the statistical comparison with other candidate models and without calculating the parameter errors. For example, without errors or confidence intervals for the exponent of the degree distribution in the nonlinear preferential attachment model, the discussion about evidence for or against a nonlinearity in the growth mechanism is not complete [5].

In addition to the methodological issues with applying the preferential attachment model to citation growth, there are also conceptual problems. The most important one regards the role of social influences on the citation dynamics. Specifically, to what extent do properties of the *authors*, e.g., their reputation as expressed by their number of previous publications or their total number of citations, play a role in citing a publication? After all, producing academic publications is a *social endeavour* that involves collaboration between authors and information spreading through social interactions.

Thus, it is reasonable to expect that these social aspects have an impact on the citation dynamics. Merton [15], in his seminal work, discussed various mechanisms of how

2470-0045/2020/102(3)/032303(14)

characteristics of authors (prominence, academic awards) may influence the recognition of their publications in terms of citations. In particular, he raised the question of whether publications by better-known researchers get more and faster recognition in the community.

To formally address such issues in a modeling approach requires us to consider not only the relations between *publications*, in a citation network, but also the relations between *authors* in a coauthorship network. Such collaboration networks between authors [16] and their coevolution with the citation networks [17] have been studied recently. For instance, it was shown that the centrality of authors in the coauthorship network prior to a publication affects the number of citations that the publication will receive [18], which allows us to predict their future success.

In this article, we provide an in-depth analysis to quantify the impact of authors' characteristics, such as their previous number of publications or coauthors, on the citation growth of their publications. Specifically, we will test different hypotheses about this impact and combine them with different growth mechanisms for the citation dynamics. This way, our results shed a new light on the discussion about the "fitness" of a publication. While it is usually attributed to the *content* of the paper, we address the question whether this "fitness" can be better attributed to the properties of the *authors*. This would allow us to transform the rather abstract notion of "fitness" in relation to "content" into a measurable and interpretable quantity.

To achieve our goal, we start from the more general perspective of coupled multilayer networks. We provide a framework to model the growth of such coupled networks, but we also demonstrate how a statistical evaluation of such growth models should be carried out. This generalizes the methods presented in Refs. [19,20] for single-layer networks. We refine these methods in two ways. First, we demonstrate how to compute the standard errors of the parameter estimates. This allows us to judge the significance of the model parameters and the corresponding model formulations. Second, we address the issue of prohibitive computational cost of the microscopic model evaluation based on the sequential addition of individual edges. We solve this for arbitrarily large networks by evaluating the models based on finite samples of growth events over the course of the network growth. By sampling events uniformly over the whole period of network growth, from the first nodes to the final observed state of the network, we ensure representing the whole growth process.

The article is organized as follows. In Sec. II we formalize (i) the multilayer network representation of citations between publications and authorship relations between publications and authors and (ii) the modeling approach to coupled growth of multilayer networks in general. In Sec. III, we introduce the maximum likelihood estimation of model parameters and model selection based on the temporal sequence of edges added to the network, the scalable estimation procedure based on event sampling, and a way to evaluate the temporal stability of model parameters. In Sec. IV we validate the methods on synthetic networks generated with known mechanisms. We then proceed with the formulation of specific models for citation growth of publications coupled with the properties of

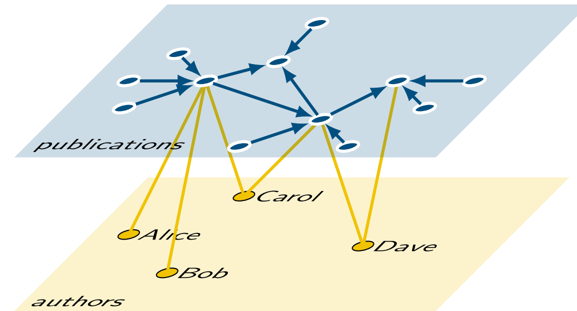

FIG. 1. Multilayer representation of authors and publications.

the authors of these publications in Sec. V. Finally, in Sec. VI, we analyze the growth of nine empirical citation networks that have been constructed from physics journals based on data from the American Physical Society and INSPIRE.

II. A MULTILAYER GROWTH MODEL

A. A multilayer approach to citation dynamics

The real-world application of our general framework comes from the domain of scientometrics. Specifically, we want to identify the mechanisms that drive citations between different publications. We therefore use data from nine physics journals, most of them from the APS, which are described in detail in Appendix C. For each journal and each publication therein, we extract data about (i) the publication time, (ii) the authors, and (iii) the publications cited in the respective publication.

To represent this relational data, we construct a two-layer network (see Fig. 1) in which the lower layer consists of the authors and the upper layer of their publications. Edges in the upper layer indicate citations between publications. The relation *between* the two layers is defined by the coauthorship for each publication. Previous works have mainly focused on the dynamics of the publication layer, which was studied in isolation, i.e., without accounting for influences from the author layer [1,2,4,7,10,21,22]. This shall be improved in our paper. The research question addressed with respect to this two-layer representation can be stated as follows: To what extent does the author layer (i.e., the status, collaboration history, etc., of authors) impact the publication layer (i.e., citation of papers)?

While this research question is interesting in itself, it does not address the notion of *time* in an explicit manner. We have to take into account that publications enter the system at different times, which restricts the range of possible citations to papers already published. With every new publication, also the author layer changes. Figure 2 illustrates with three time-ordered snapshots such a coupled growth process. To capture this dynamics properly, in this paper we formulate a *growth* model for the publication layer that takes changes in the author layer into account. The related research question can be stated as follows: What is the impact of time on these relations, specifically the decay of attention toward older publications?

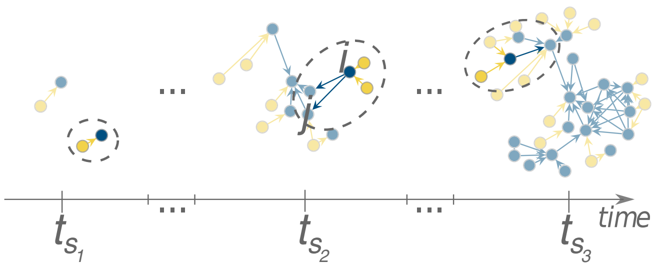

FIG. 2. Coupled dynamics of the two-layer network. Blue nodes or edges refer to the publication layer and yellow nodes or edges to the author layer (see also Fig. 1). Each growth event is associated with a new publication, which draws citation edges to older publications and coauthorship edges between the corresponding authors.

B. A multilayer approach to coupled growth processes

The application of multilayer networks to scientometrics is just one instance of a bigger class of problems, in which two processes, *within* and *across* different layers, concurrently determine the overall dynamics. Therefore, in this paper we want to derive a general approach to these problems before applying it to the specific application of the citation dynamics.

We consider a growing multilayer network $G(t) = G[V(t), E(t)]$ with L sets of nodes $V(t) = \{V^{\mu}(t)\}_{\mu\in[1,L]}$ that define the network layers $\mu \in [1, L]$ and with edges $E(t) = \{E^{\mu\nu}(t)\}_{(\mu,\nu)\in[1,L]\times[1,L]}$. Each set of edges $E^{\mu\nu}(t)$ comprises all and only edges that connect nodes in layer $V^{\mu}(t)$ to nodes in $V^{\nu}(t)$. The left panel in Fig. 3 illustrates such a multilayer network where the edges within and between two layers μ and ν shown. In principle, the edges within and across different layers can be directed or undirected independently from each other. Edges can also be weighted, but for simplicity we will not consider the weights in the further discussion. This notation above is quite general and can encode the special cases of multiplex networks [23] in which $V^{\mu}(t) = V^{\nu}(t)$ for all $\mu, \nu \in [1, L]$ or multipartite networks [24] in which $E^{\mu\mu} = \emptyset$ for any μ and ν.

In the following we investigate *growing* networks, i.e., nodes and edges are added to the network over time. We do not consider isolated nodes and the removal of nodes or edges. Formally, our models define the probability of adding an *edge* $e_{ij} \in E^{\mu\nu}(t+1)$ to the network at time $t+1$. We assume discrete time, parametrized by the network growth. This means that a time step $t \to t+1$ occurs only when at least one edge (but potentially, multiple simultaneous edges) is added to the network. The probability of this event is, in its most general form, expressed as:

$$\Pr[e_{ij} \in E^{\mu\nu}(t+1) \wedge e_{ij} \notin E^{\mu\nu}(t)] = \pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]. \quad (1)$$

$\pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]$, which is the central element of our model specifications below, defines this probability as a function of the network $G(t)$, taken at the time before the edge is added, and a vector of parameters, $\boldsymbol{\theta}$, which later allows us to distinguish between different models.

Note that, in the right panel of Fig. 3, the $\pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]$ denote *model components*. They may differ for different sets of edges and are therefore indexed by the network layers μ and ν. In principle, each set of edges $E^{\mu\nu}$ can be modelled by its own model component. Because each model component is in general a function of the whole network at time t, the addition of edges in one network layer is affected by the state of the network not only in that layer but also in other layers. In other words, this general formulation allows for coupling in the growth dynamics of different layers. In Sec. IV, we will specify the $\pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]$ to capture the growth of a synthetic network. More importantly, in Sec. V A we will present a general expression for the model component $\pi_{ij}^{pp}[G(t); \boldsymbol{\theta}]$ that shall describe the growth of citations between publications. There we make use of five parameters $\boldsymbol{\theta} = [\alpha, \beta, \gamma, \delta, \tau]$ that will allow us to include or exclude different mechanisms that impact citations. Their importance can be tested in a statistical approach described in the following section.

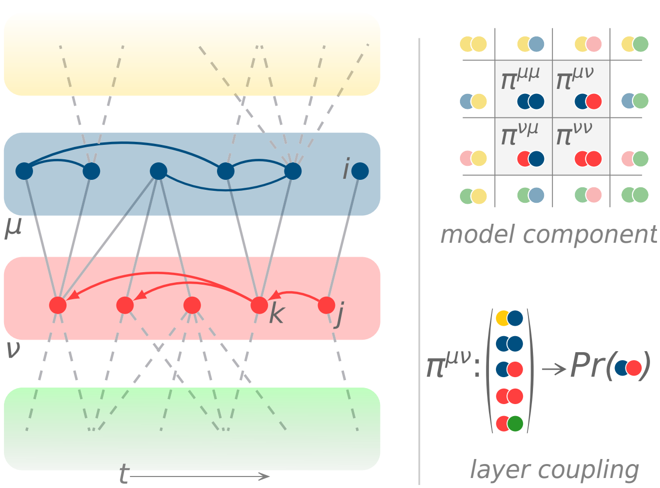

FIG. 3. Left: Multilayer network representation. For illustrative purposes, only edges between neighboring layers are drawn. Right top: The growth of the network can be described by means of separate model components $\pi^{\mu\nu}$, each corresponding to edge creation of one type, $E^{\mu\nu}$. Right bottom: The coupling between the dynamics of different layers is captured in the model components.

III. TIME-RESPECTING MAXIMUM LIKELIHOOD ESTIMATION

A. Estimation of model parameters

Having defined multilayer network growth by means of model components $\pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]$, here we describe a maximum likelihood estimation of the model parameter vector $\boldsymbol{\theta}$, given a growing network $G(t)$. Similarly to the approach in Refs. [19,20], we calculate the model likelihood under the assumption of statistical independence of *adding* individual edges:

$$\ln \mathcal{L}[\boldsymbol{\theta}; G(T)] = \sum_{t=1}^{T-1} \sum_{\mu,\nu} \sum_{\substack{e_{ij} \in E^{\mu\nu}(t+1) \\ e_{ij} \notin E^{\mu\nu}(t)}} \ln \pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}], \quad (2)$$

where T is the last observed time. Note that the independence assumption does not apply to the edges themselves: The probability of an edge depends on the network state at time

t and thus on the previously existing edges. But adding one edge at a given time does not depend on other edges added at the same time. The vector of model parameters $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood of the model given the network is found by solving the equation

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = 0. \tag{3}$$

Equation (3) allows us to find the best estimate of the given model's parameter vector. However, it does not provide information about the *uncertainty* of the estimates. Fortunately, the latter can also be computed as part of the maximum likelihood estimation (MLE) method. It is an important property of the MLE method that, under some mild conditions, the vector of parameter estimates $\hat{\theta}$ follows a multivariate normal distribution around the true value of the parameters. This means that the uncertainty about the parameter estimates can be adequately described using standard errors based on the corresponding covariance matrix. In the context of MLE, the covariance matrix of the parameters is then given by the inverse of the *Fisher information matrix* $\mathcal{I}(\theta)$:

$$\sigma^2(\hat{\boldsymbol{\theta}}) = [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}. \tag{4}$$

The matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ is the expectation of the second derivatives of the log-likelihood function, which is usually not feasible to find. Hence, in practice it is commonly approximated by the *observed* Fisher information matrix:

$$\mathcal{I}(\boldsymbol{\theta}) \approx \mathcal{I}(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}, \tag{5}$$

i.e., $\mathcal{I}(\boldsymbol{\theta})$ corresponds to the values of the second derivatives as in Eq. (5) calculated for the parameter estimate $\hat{\boldsymbol{\theta}}$. In general, the MLE of the parameters $\boldsymbol{\theta}$ is not performed by finding the maxima of Eq. (3) analytically but is computed by means of numerical optimization as discussed below in Sec. III C.

B. Model selection

So far, we have specified our growth model by means of model components $\pi_{ij}^{\mu\nu}[G(t);\boldsymbol{\theta}]$, where the model parameter vector $\boldsymbol{\theta}$ was determined by a MLE. Model selection denotes a defined procedure of statistically comparing different models in their ability to capture the data given with the growing network $G(t)$. With "different models," we refer not only to different parameter vectors $\boldsymbol{\theta}$ but also to different formulations of sets of model components $\pi_{ij}^{\mu\nu}$. For instance, we can define a model A that implements information from two sets of edges, $E^{\mu\nu}$ and $E^{\mu'\nu'}$, and then compare it to a model B based on only one set of edges. In general, model A is expected to describe the data better, in statistical terms, because it takes more data into account. Therefore, model selection has to consider not only the goodness of a fit but also the *model complexity* as expressed by the degrees of freedom of the model: The more information we put into the model, the more complex it is.

Later, in Sec. VI, we will define seven models for the growth of citations among scientific publications. In some of these models the new citations will depend only on the previous citations, while in other models the new citations will also depend on authorship relations. To compare these models, we will compute their *relative likelihoods*, similarly to [20]:

$$w_m = \exp\left\{\tfrac{1}{2}[\min_m(\mathrm{AIC}_m) - \mathrm{AIC}_m]\right\}, \tag{6}$$

where AIC_m is the value of Akaike information criterion of the model m. It is based on the maximum likelihood value of the model, corrected for the model complexity:

$$\mathrm{AIC}_m = -2\log \mathcal{L}[\hat{\boldsymbol{\theta}}_m; G(T)] + 2|\boldsymbol{\theta}_m|, \tag{7}$$

where $|\boldsymbol{\theta}_m|$ denotes the size of vector $\boldsymbol{\theta}_m$, i.e., the number of parameters of the model m. For better interpretability, after computing the values of w_m according to Eq. (6), they are normalized to add to 1. Our choice of the relative likelihoods w_m of models is motivated by their applicability for comparing a wide range of models. This is in contrast with, e.g., likelihood ratio tests that could be used only for comparing so-called nested models where one model is a special case of another.

C. Scalable MLE by event sampling

While the MLE cannot be solved *analytically*, we can make use of the fact that, for commonly studied network growth models, the likelihood function has a convex shape (see Appendix A and [20]). This means that the Maximum Likelihood Estimation (MLE) can be done *numerically* by using a greedy hill-climbing algorithm. In the following, we will use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. It has the advantage that, while finding the function extremum, it also estimates the inverse of the Hessian matrix, which for the log-likelihood function corresponds to the Fisher information matrix. This means that we obtain both the MLE of the parameters and their variances at the same time [see Eq. (5)]. More specifically, to obtain our results we use a variant of the algorithm, `L-BFGS-B`. It allows to set boundary constraints on the parameters and uses limited memory, making the whole procedure more scalable [25,26]. `L-BFGS-B` is a standard algorithm for solving nonlinear optimization problems with bounded variables and implementations are available in common scientific computing libraries [27]. For further details on the algorithm's complexity, we refer to the original publications [25,26].

Performing a MLE for Eq. (2) based on the full network data becomes very costly for large networks. For each time step the computations for $\pi_{ij}^{\mu\nu}[G(t);\boldsymbol{\theta}]$ have to be redone based on the network state at that time. Therefore, the total computation time scales at least with the square of the final network size. The larger the network, the more distinct time steps we have to consider to describe its growth (as time is parametrized by edges added, see Sec. II B). To reduce the computational cost of the MLE, we apply the model evaluation of Eq. (2) only to a *subset* of growth events observed over time $t \in [1, T]$. For this, we sample, without replacement, a number $S < T$ of time steps uniformly at random, $\{t_s\}_{s\in[1,S]}$. This set of time stamps uniformly covers the full time range $[1, T]$ of the network growth, provided that the largest sampled time step t_S is close to the final time T. Then we can compute the MLE of a model based on this sample of time steps:

$$\ln \mathcal{L}(\boldsymbol{\theta}; G(T), S) = \sum_{s=1}^{S} \sum_{\mu,\nu} \sum_{\substack{e_{ij} \in E^{\mu\nu}(t_s+1) \\ e_{ij} \notin E^{\mu\nu}(t_s)}} \ln \pi_{ij}^{\mu\nu}[G(t_s); \boldsymbol{\theta}]. \tag{8}$$

Through the model components $\pi_{ij}^{\mu\nu}[G(t_s);\boldsymbol{\theta}]$ potentially all nodes and edges present in $G(T)$ will affect the log-likelihood of the model given by Eq. (8). In other words, the likelihood estimation is based on "probes" of temporal events over the whole growth period of the network. By fixing the number of time steps, S, the computation time can, for the final network size, be decreased by an order of magnitude. As a trade-off, the precision of the parameter estimates will be lower because their variance is inversely proportional to the number of observations used in the estimation. This trade-off can be controlled by adjusting the number S of sampled time stamps according to the desired precision of the parameter estimate.

D. Temporal stability of a model

The above discussion assumes that the growth model does not *explicitly* depend on time, even though the studied network changes over time. In particular, the parameter vector $\boldsymbol{\theta}$ used to define the model components $\pi_{ij}^{\mu\nu}[G(t_s);\boldsymbol{\theta}]$ does not depend on time. To confirm that this assumption holds we split the whole growth time period $[1,T]$ into into R consecutive time windows such that each of them contains approximately $\lfloor T/R \rfloor$ time steps. We then partition the log-likelihood function, Eq. (2), into the sum of log-likelihoods computed for each time window separately:

$$\ln \mathcal{L}[\boldsymbol{\theta}; G(T)] = \sum_{r=1}^{R} \ln \mathcal{L}[\boldsymbol{\theta}; G, r], \tag{9}$$

where each summand on the right-hand side is given by

$$\ln \mathcal{L}[\boldsymbol{\theta}; G, r] = \sum_{t=(r-1)\lfloor \frac{T}{R} \rfloor}^{r\lfloor \frac{T}{R} \rfloor} \sum_{\mu,\nu} \sum_{\substack{e_{ij} \in E^{\mu\nu}(t+1) \\ e_{ij} \notin E^{\mu\nu}(t)}} \ln \pi_{ij}^{\mu\nu}[G(t); \boldsymbol{\theta}]. \tag{10}$$

Equation (10) provides the log-likelihood of the model computed for a certain time window $r \in [1,R]$. Based on this, we can estimate the model parameters for each time window separately using the MLE described earlier. The assumption of a stationary model that itself does not depend on time is tested by comparing the parameter estimates for the R consecutive time windows. We will use this approach in the next section to validate the method based on synthetically generated networks with known growth mechanisms.

IV. METHOD VALIDATION ON SYNTHETIC DATA

To confirm that the presented method leads to correct results, we first use a synthetic network comprising two sets of nodes denoted as V^p and V^a and two sets of directed edges E^{pp} and E^{ap}. That is, we have a two-layer network with edges within layer p and between the two layers, p and a. The set V^a consists of a fixed number of initially disconnected nodes. During the growth of the network, these nodes will be subsequently connected to the nodes in layer p, as described below.

Hence, for the growth of the synthetic network we concentrate on layer p, where one node i at a time is added with an *out-degree* randomly drawn between one and five. This node connects to already existing nodes based on *linear preferential attachment*, i.e., on the *in-degrees* of the chosen nodes. Because of the two-layer network, nodes are connected *within* and *across* layers. Consequently, the first growth mechanism takes the in-degree of nodes *within* a layer, i.e., the edges in the set E^{pp}, into account. The second growth mechanism, on the other hand, builds on the in-degree of nodes *across* layers, i.e., the edges in the set E^{ap}.

Using the combined information from the two layers, we specify the growth mechanism as follows: When adding an edge from the newly added node $i \in V^p$, we choose the existing node $j \in V^p$ with probability:

$$\pi_{ij}^{pp,\text{true}}[G(t);\alpha,\delta] = \alpha \frac{k_j^{pp,\text{in}}(t)+\delta}{\sum_{l\in V^p(t)}\left[k_l^{pp,\text{in}}(t)+\delta\right]} + (1-\alpha)\frac{k_j^{ap}}{\sum_{l\in V^p(t)} k_l^{ap}}, \tag{11}$$

where $k_j^{pp,\text{in}}(t)$ is the number of incoming edges of node $j \in V^p$ within layer p, i.e., its in-degree with respect to the edge set $E^{pp}(t)$, and $k_j^{ap}(t)$ is the number of edges of node j to nodes in layer a, i.e., its in-degree with respect to the edge set $E^{ap}(t)$. The parameter $\alpha \in [0,1]$ weights between the two growth mechanisms (often called mixture weight).

The constant δ [8,21], which appears in the first growth mechanism, is added to the in-degree of nodes in layer p to ensure that, in the case of directed networks, a newly added node with nonzero out-degree but *zero in-degree* can attract incoming edges. For the second mechanism, we do not need to include such an additive constant because each node in $i \in V^p$ has at least one edge $e_{ij} \in E^{ap}$ connecting it to a node $j \in V^a$ in the second layer. Specifically, we connect the newly added node $i \in V^p(t)$ to between one and five nodes in layer a with a uniform probability:

$$\pi_{ji}^{ap,\text{true}}[G(t)] = \frac{1}{|V^a(t)|}. \tag{12}$$

For the experiment below, we choose $V^p(T) = V^a(T) = 2000$ nodes in each of the two layers. Adding one node at a time to layer p, this implies a total growth period of $T = 2000$. To test whether our procedure works, we fix the values of the parameters as $\delta = 1.2$ and $\alpha = 0.7$. Once we generate a synthetic network, we can apply the MLE to recover the (known) parameters of the model. Given the sampling procedure described in Sec. III C, we estimate the parameters for varying sizes $S \leqslant T$ of the randomly sampled time stamps. Additionally, we repeat the procedure 10 times for each sample size.

The results of this numerical experiment are reported in Fig. 4. We find that (i) the estimates are close to the true value and the true value falls within one standard error of the estimates and (ii) the variance decreases with the sample

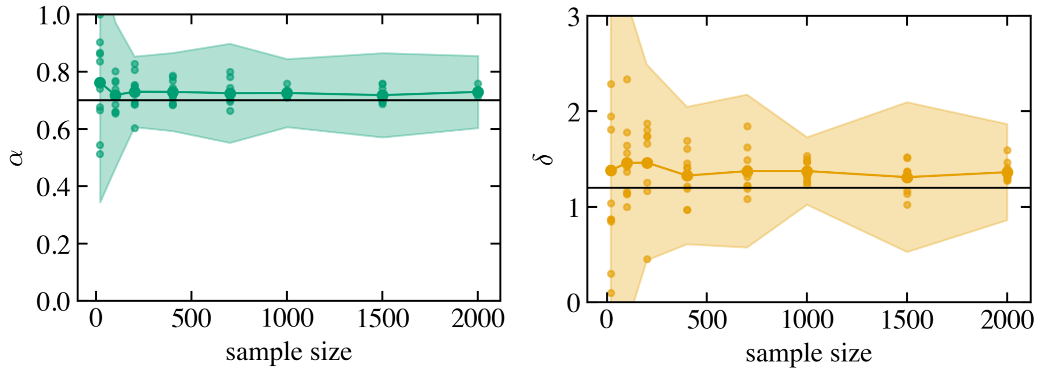

FIG. 4. Parameter estimates (small dots) of the model given in Eq. (11) for varying sample size, with 10 realizations for each sample size. These estimates (large dots and lines) and their standard errors (shaded area) are pooled for each sample size. The black line corresponds to the true parameter value used for generating the networks.

size as expected. The decrease becomes noticable beyond a sample size of about 400, i.e., 20% from the full data. The results confirm that our method is able to identify the correct parameters.

Next we want to test whether the parameter estimates are stable over different growth periods of the network, as described in Sec. III D. For this test, we partition the total growth period of $T = 2000$ time stamps into $R = 10$ time windows. We then estimate the model parameters for each time window based on the log-likelihood defined in Eq. (10). The results for five synthetically generated networks are shown in Fig. 5. We see that for both parameters the true value lies within one standard error of the estimates for each time window. Further, no trend is observed in the parameter estimates over time. This confirms that our method is robust also with respect to the sampling procedure.

Last, it is important to also demonstrate that the *model selection approach* based on the relative likelihoods (Sec. III B) can correctly identify the best model among various candidates. To this end, we eventually evaluated a set of different models (explained in the next section) in addition to the known true model of Eq. (11). We show these results in Table III. They confirm that, indeed, the model selected this way corresponds to the true model used to generate the synthetic data set, i.e., that the correct model was identified.

V. MODEL SPECIFICATION FOR CITATION DYNAMICS

Now that we have seen how our modeling framework can be used and how it is able to correctly identify the mechanisms underlying a synthetic growth process, we are ready to specify the framework for the real-world application in scientometrics. We recall that model specification implies to determine the model components $\pi_{ij}^{\mu\nu}[G(t);\boldsymbol{\theta}]$ from Eq. (1). Specifically, to model the two-layer network with publications V^p and authors V^a we need to determine the four model components shown in Fig. 3. This requires us to come up with some dynamic hypotheses of how these two layers are coupled, which should be aligned to the empirical facts.

A. The model component π_{ij}^{pp}

Given the dynamics for our network illustrated in Fig. 2, we can make two observations. First, authors do not have direct links but only via coauthored publications. Therefore, the probability $\pi_{ij}^{aa}[\cdot]$ does not need to be modelled. Second, relations between authors and papers are undirected, and thus $\pi_{ij}^{ap}[\cdot] = \pi_{ij}^{pa}[\cdot]$ holds. Furthermore, we assume that the links between authors and publications are given by the data,

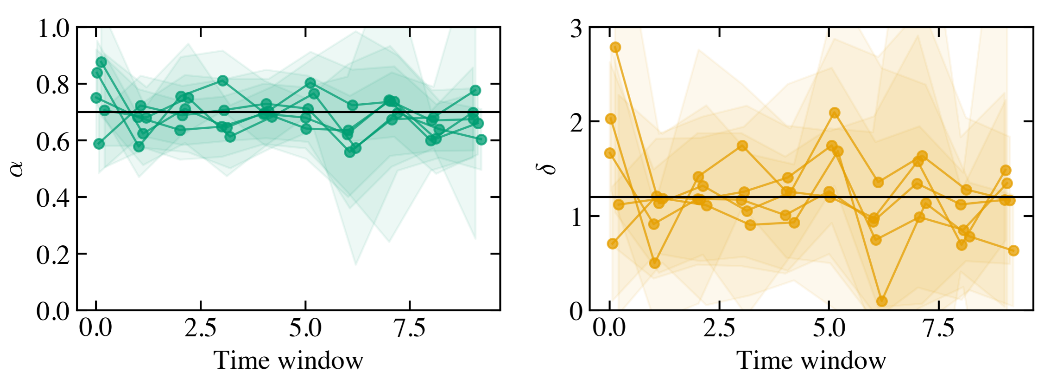

FIG. 5. The observed growth of a network is partitioned into 10 consecutive time windows, each comprising approximately the same number of time steps. The model in Eq. (11) is then evaluated based on each time window separately. The resulting parameter estimates (dots and lines) and their standard errors (shaded areas) are shown for five synthetic networks. The black line corresponds to the true parameter value.

making $\pi_{ij}^{ap}[\cdot]$ deterministic. Therefore, we are only left with specifying a single model component, $\pi_{ij}^{pp}[\cdot]$. This defines the probability of an incoming publication i to form an edge, i.e., to cite, an existing publication j. In most general terms, this probability can be written in full as:

$$\pi_{ij}^{pp}[G(t);\boldsymbol{\theta}] = \alpha \frac{\eta_j[G(t)]^\beta \times K_j[G(t);\gamma,\delta,\tau]}{\sum_{l\in V^p(t)} \eta_l[G(t)]^\beta \times K_l[G(t);\gamma,\delta,\tau]} + (1-\alpha)\frac{\eta_j[G(t)]}{\sum_{l\in V^p(t)} \eta_l[G(t)]}, \tag{13}$$

$$\boldsymbol{\theta} = [\alpha,\ \beta,\ \gamma,\ \delta,\ \tau]', \tag{14}$$

which follows the structure already discussed for Eq. (11). Note that, once published, the citations in publication are generally not further modified. Hence, we do not model the probability of existing publications to form new links. Equation (13) comprises two constituents, $K_j[\cdot]$ and $\eta_j[\cdot]$: With $K_j[\cdot]$, we specify the *citation self-dynamics*, in particular different specifications from the preferential attachment family. $\eta_j[\cdot]$ shall consider the *fitness* of the publication. It describes, for instance, how attractive the publication is for citations [10] and relates to the fitness term in other preferential attachment models [9,28]. We want $\eta_j[\cdot]$ to capture *noncitation related influences*, specifically influences from the authors of the paper on its probability to receive citations. Therefore, we will refer to $\eta_j[\cdot]$ as the *social constituent*. Both components will be individually addressed in the following sections.

The parameters $\alpha \in [0, 1]$ and β can be used to couple the citation and the social constituents either additively or multiplicatively. For $\beta = 0$, we obtain

$$\pi_{ij}^{pp}[G(t);\boldsymbol{\theta}] = \alpha \frac{K_j[G(t);\gamma,\delta,\tau]}{\sum_{l\in V^p(t)} K_l[G(t);\gamma,\delta,\tau])} + (1-\alpha)\frac{\eta_j[G(t)]}{\sum_{l\in V^p(t)} \eta_l[G(t)]}, \tag{15}$$

which means the probability $\pi_{ij}^{pp}[\cdot]$ for publication i to cite publication j is the weighted mixture of the two constituents. Such *mixture models* are used when the statistical population is known to comprise subpopulations that are described by different probability distributions. The mixture weight α is, then, determined by the relative size of each subpopulation. In our case, it means that node j is chosen *either* based on constituent $K_j[\cdot]$ of citation self-dynamics *or* based on the social constituent $\eta_j[\cdot]$, with the respective weight α.

For $\alpha = 1$, we obtain multiplicative coupling:

$$\pi_{ij}^{pp}[G(t);\boldsymbol{\theta}] = \frac{\eta_j[G(t)]^\beta \times K_j[G(t);\gamma,\delta,\tau]}{\sum_{l\in V^p(t)} \eta_l[G(t)]^\beta \times K_l[G(t);\gamma,\delta,\tau]}. \tag{16}$$

It means that the social constituent *scales* the effect of the citation constituent during the growth process, with $\beta \in [0,\infty)$ as the *impact* of the social constituent. Hence, the growth mechanism defined within the citation layer is the main, *baseline*, effect while the social constituent biases it. For intermediate values of α and β mixtures between additive and multiplicative coupling can be achieved.

B. Citation constituent

To define the citation constituent $K_j[\cdot]$, we use a general expression for preferential attachment:

$$K_j[G(t);\boldsymbol{\theta}] = \left[k_j^{pp,\mathrm{in}}(t) + \delta\right]^\gamma e^{-\frac{t-t_j}{\tau}}. \tag{17}$$

The term $k_j^{pp,\mathrm{in}}(t) + \delta$ describes linear preferential attachment, where $k_j^{pp,\mathrm{in}}(t)$ denotes the in-degree of the publication j in the citation layer at time t, i.e., before citations made to j are considered. The constant δ, again, corrects for publications without citations (i.e., in-degree zero) at time t [29]. The parameter $\gamma \in [0,\infty)$ allows us to study *nonlinear preferential attachment* by controlling the degree-related preference. Specifically, it amplifies or weakens the effect of the number of previous citations. Finally, the term $e^{-(t-t_j)/\tau}$ introduces a *relevance decay* [13]. Here t_j denotes the time at which the publication j was added to the network, thus the probability for a publication to be cited now also depends on its age. For simplicity, we will measure time using the number of publications in layer p. This means $t_j = j$ if nodes are labeled in the order they were added to the network. The parameter $\tau \in [0,\infty)$ determines the characteristic time of the relevance decay.

In our empirical analysis, we study four specifications of the citation constituent $K_j[\cdot]$ defined by certain settings of the parameters γ, δ, and τ as follows:

(i) Uniform Attachment (UNIF): This model defines a baseline without citation preferences for any paper. This is achieved, e.g., with $\gamma = \delta = 0$ and $\tau = \infty$, yielding

$$K_j[G(t);\boldsymbol{\theta}] = 1. \tag{18}$$

(ii) Preferential Attachment (PA): This model is motivated by the general finding that highly cited publications tend to receive more citation in the future [8]. This can be described by linear preferential attachment, which is obtained for $\gamma = 1$ and $\tau = \infty$,

$$K_j[G(t);\boldsymbol{\theta}] = k_j^{pp,\mathrm{in}}(t) + \delta. \tag{19}$$

(iii) Preferential Attachment with Relevance Decay (PA-RD): The motivation for this model is similar to PA, but it additionally considers that papers are cited less over time [13], for which the most evidence can be found in literature. This is obtained for $\gamma = 1$,

$$K_j[G(t);\boldsymbol{\theta}] = \left[k_j^{pp,\mathrm{in}}(t) + \delta\right] e^{-\frac{t-t_j}{\tau}}. \tag{20}$$

(iv) Nonlinear Preferential Attachment with Relevance Decay (PA-NL-RD): The added nonlinearity to the PA-RD model allows us to scale the citation probability with the current citation count [5,29]. To keep the same number of degrees of freedom for this model compared to PA-RD, we fix the offset $\delta = 1$, yielding

$$K_j[G(t);\boldsymbol{\theta}] = \left[k_j^{pp,\mathrm{in}}(t) + 1\right]^\gamma e^{-\frac{t-t_j}{\tau}}. \tag{21}$$

C. Social constituent

To completely specify the model component $\pi_{ij}^{pp}[\cdot]$, Eq. (13), we have to make assumptions about the social constituent, $\eta_j[G(t)]$. Here we mostly follow the arguments made by Merton in his seminal paper [15], where he theorizes that the academic prominence of authors influences the success and recognition of their future work. To quantify different levels of prominence, we use information related to the number of coauthors and the number of their previous publications, as specified in the following:

(i) Number of Authors (NAUT): As the simplest possibility, we define the social constituent as the number of authors that wrote the publication. We argue that more coauthors (i) increase a paper's potential social exposure and (ii) allow for a better division of labor and to a higher impact [30]. A paper with a higher number of coauthors should therefore receive more citations. Mathematically, this is formulated as

$$\eta_j[G(t)] = k_j^{ap}(t), \tag{22}$$

where $k_j^{ap}(t)$ counts the number of authors of a publication.

(ii) Number of Previous Co-authors (NCOAUT): We assume that authors who have previously collaborated with a larger number of coauthors have more experience and more visibility in their community, which in turn increases the citations for their new publications. To calculate, for the coauthors of a given paper, their set of unique *previous* coauthors, we use:

$$\eta_j[G(t)] = \left|\left\{v \mid \exists\lambda_{3,jv}^{ap}(t)\right\}\right|, \tag{23}$$

That means that we count the unique endpoints v of the self-avoiding paths $\lambda_{3,jv}^{ap}$ of length three over author-publication edges, i.e., paths `publication` → `authors` → `previous publications` → `previous coauthors`.

(iii) Maximum Number of Previous Co-authors (MAXCOAUT): In this variant of `NCOAUT` we only take the author with the largest number of previous coauthors into account [18]. This is motivated by Merton [15], who argues that the academic credit for a publication written by a team of authors is often attributed to the most prominent author. For authors $q \in V^a(t)$ of the publication j this yields

$$\eta_j(t) = \max_q \left(\left|\left\{v \mid \exists\lambda_{2,qv}^{ap}(t)\right\}\right|\right). \tag{24}$$

This means that we count the unique endpoints v of the self-avoiding paths $\lambda_{2,jv}^{ap}$ of length two: `author` → `previous publications` → `previous coauthors`.

(iv) Number of Previous Publications (NPUB): We argue that with more prior publications, i.e., with increasing experience of an author, (i) the impact of new publications and (ii) the author's recognition in the academic community should both increase, which leads to increased citation counts of future publications. Hence, we consider the number of distinct publications written by all authors of publication j,

$$\eta_j(t) = \left|\left\{v \mid \exists\lambda_{2,jv}^{ap}(t)\right\}\right|. \tag{25}$$

Similarly to Eqs. (23) and (24), we count the unique endpoints v of the self-avoiding paths $\lambda_{2,jv}^{ap}$ of length two: `publication` → `authors` → `other publications`.

(v) Maximum Number of Previous Publications (MAXPUB): In this variant of `NPUB` we assume that the probability to be cited only depends on the maximum number of previous publications among all coauthors. For authors $q \in V^a(t)$ of the publication j, this is formulated as

$$\eta_j(t) = \max_q \left(k_q^{ap}(t)\right), \tag{26}$$

where k_q^{ap} is the degree of author q with respect to author-publication edges at time t.

Based on the three alternatives for the citation constituent and the five alternatives for the social constituent, we can now create a large number of possible models, i.e., expressions for $\pi_{ij}^{pp}[\cdot]$, Eq. (13), by using either additive or multiplicative coupling of the constituents. Three models (`UNIF`, `PA-RD`, and `PA-NL-RD`) consider only the citation self-dynamics, i.e., there is no coupling to the social constituent. Because we take from the literature that `PA-RD` is the most promising candidate for the citation constituent, we restrict the combined models only to combinations of `PA-RD`. For the additive combination, we test it together with all five alternatives for the social constituent for the multiplicative combination only with the two most promising ones. A summary of all tested models and their free parameters is given in Table I.

VI. EMPIRICAL AUTHORSHIP-CITATION NETWORKS

The bibliographic data used for our model testing come from two sources: APS journals [31] and INSPIRE [32]. The details of these data, their origin and our motivation to use them, are described in the Appendix C. Here we only list the nine different journals that were included in our study:

Physical Review (PR)
Physical Review A (PRA)
Physical Review C (PRC)
Physical Review E (PRE)
Reviews of Modern Physics (RMP)
The Journal of High Energy Physics (JHEP)
High Energy Physics in Physical Review Journals (PR-HEP)
Physics Letters A and *B* (Phys. Lett.)
Nuclear Physics (Nuc. Phys.)

A. Results for nine empirical citation networks

In the following, we fit, for each of the nine citation networks listed above, our 11 model specifications for $\pi_{ij}^{pp}[G(t);\boldsymbol{\theta}]$ presented in Table I. Subsequently, we apply the model selection technique described in Sec. III to select the most adequate model.

One could perform a stepwise model selection as follows: For each two-layer network, first fit the models with only the citation constituent and choose the one with the highest relative likelihood. Then, consider models with couplings between citation and social constituents based on the previously selected citation constituent. Instead, we chose the fixed set of models with and without coupling based on our prior beliefs. Namely, we hypothesize that (i) the most appropriate citation constituent is the the linear preferential attachment with relevance decay, `PA-RD`, and (ii) the additive

TABLE I. Specification and relationships of the seven models used in the analysis. Columns with ? denote the free parameters that are fitted for the model. PA: plain preferential attachment; PA-RD: preferential attachment with relevance decay; PA-NL-RD: preferential attachment with with relevance decay and added nonlinearity. In addition we consider models that combine PA-RD with different candidates for the social constituent (see Sec. V C for explanation). All models are compared to a baseline model with uniform attachment (UNIF).

Model	**Citation constituent**	**Social constituent**	α	β	γ	δ	τ
UNIF	UNIF	—	1	0	0	0	∞
PA	PA	—	1	0	1	**?**	∞
PA-RD	PA-RD	—	1	0	1	**?**	**?**
PA-NL-RD	PA-NL-RD	—	1	0	**?**	1	**?**
PA-RD + NAUT	PA-RD	NAUT	**?**	0	1	**?**	**?**
PA-RD + NCOAUT	PA-RD	NCOAUT	**?**	0	1	**?**	**?**
PA-RD + MAXCOAUT	PA-RD	MAXCOAUT	**?**	0	1	**?**	**?**
PA-RD + NPUB	PA-RD	NPUB	**?**	0	1	**?**	**?**
PA-RD + MAXPUB	PA-RD	MAXPUB	**?**	0	1	**?**	**?**
PA-RD × NCOAUT	PA-RD	NCOAUTH	1	**?**	1	**?**	**?**
PA-RD × NPUB	PA-RD	NPUB	1	**?**	1	**?**	**?**

coupling between the constituents is more appropriate than the multiplicative one. We will be able to judge whether there is strong evidence against these hypotheses based on the results of model selection. For instance, if the more complex nonlinear preferential attachment in the citation constituent is more prevalent, then `PA-NL-RD` will be selected as the best model (as we will see, this is the case for some networks). Or we would find very large values of the characteristic relevance decay time τ (of the order of, or larger than, the whole growth time) if the simplest preferential attachment without relevance decay would be a better choice for the citation constituent in a coupled model.

We proceed as follows: For each network, we estimate the parameters based on 5000 publications, sampled uniformly at random, which are added to the network during its growth, as explained in Sec. III C. The exception is Reviews of Modern Physics (RMP), which has in total only 3006 publications, so we considered all of them for model fitting. The number of citation edges $|E^{pp}|_S$ corresponding to the sample publications and used in likelihood calculation of Eq. (8) varies between 4318 for RMP and 60131 for The Journal of High Energy Physics (JHEP).

In Table II, we present the selected models that have a relative likelihood $w_m \geqslant 0.01$ for each of the networks. The table shows the estimates of the free parameters of the corresponding model, its relative likelihood w_m within the pool of the selected candidate models and the log-likelihood of the model per considered edge. We chose the latter instead of the total log-likelihood or the AIC score, because these strongly depend on the number of edges which differs across the studied networks. Even though we perform the likelihood estimation based on a fixed sample size of 5000 publications, the density of the networks differs. Note that for RMP, three different models are listed because the model selection based on relative likelihood was inconclusive. However, in all of these three models the social constituent accounting for authors' number of publications is present. The inconclusive results for RMP are likely due to the smaller size of RMP compared to the other networks. For all other journals the data provided enough evidence to select one model over the other candidates.

Before discussing these results, we confirm that the model parameters are stable over the whole growth period of a network. We illustrate this using the example of JHEP. Following

TABLE II. Parameter estimations for the selected growth models of nine physics journals. For the fixed parameters of the different models see Table I.

Journal	**Selected models**	$\ln \mathcal{L}/\lvert E^{pp}\rvert_S$	w_m	α	β	γ	δ	τ
PR	PA-RD + NCOAUT	−4.00314	1.00	0.90(±0.020)	—	—	0.95(±0.247)	5185
PRA	PA-NL-RD	−4.13827	1.00	—	—	1.14(±0.224)	—	8411
PRC	PA-NL-RD	−3.92745	1.00	—	—	1.09(±0.007)	—	4860
PRE	PA-NL-RD	−4.14851	1.00	—	—	1.21(±0.013)	—	9706
	PA-RD + NPUB	−2.84885	0.71	0.85(±0.124)	—	—	0.47(±0.086)	480
RMP	PA-RD + MAXPUB	−2.84909	0.26	0.84(±0.016)	—	—	0.45(±0.032)	479
	PA-RD × NPUB	−2.84959	0.03	—	0.19(±0.021)	—	0.74(±0.315)	480
JHEP	PA-RD + NCOAUT	−3.60015	1.00	0.85(±0.007)	—	—	1.11(±0.074)	2828
PR-HEP	PA-RD + NPUB	−3.99631	1.00	0.87(±0.265)	—	—	0.65(±1.403)	7347
Phys. Lett.	PA-RD + NPUB	−3.56096	1.00	0.81(±0.075)	—	—	0.40(±0.186)	2991
Nuc. Phys.	PA-RD + NPUB	−3.61780	1.00	0.85(±0.129)	—	—	0.51(±2.072)	3278

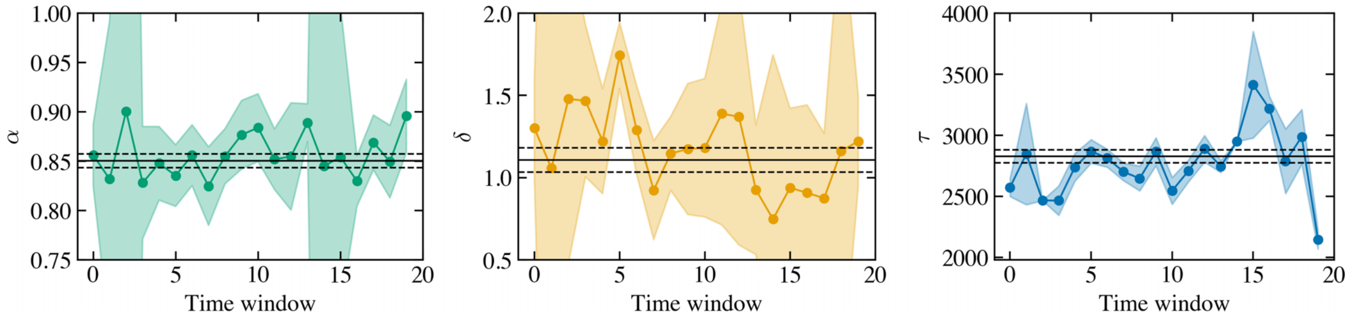

FIG. 6. Model parameters of PA-RD-NCOAUT estimated for JHEP from 5000 sampled publications. The solid black lines correspond to the parameters and the region between the dashed lines to their standard errors estimated on all 5000 sampled publications. The whole growth period (x axis) was divided into 20 consecutive time windows, each comprising approximately 250 sampled publications. The colored dots and lines indicate the parameters and the shaded area their standard errors within each time window.

the procedure described in Sec. III D, we first partition the whole growth period into 20 consecutive time windows and then estimate the parameters for the selected best model for this network, PA-RD-NCOAUT, based on each time window separately. The results for the three parameters of the model are shown in Fig. 6.

One insight regards the relation between the estimates α^T, δ^T, τ^T based on the whole sample of 5000 publications (black lines) and the estimates α^R, δ^R, τ^R based on the much-smaller subsamples for each time window. For the parameter δ, we see that the standard error for δ^T is within the standard errors for δ^R for all but one estimates. Similarly, the standard error for α^T is within the standard errors for α^R for 18 of 20 estimates. Hence, for these two parameters we can conclude that the MLE leads to a feasible outcome and that the parameters are stable over time.

For the third parameter, τ, we also do not see any temporal trend in the estimates. However, the standard errors for τ^R are unexpectedly small and their magnitude becomes comparable to τ^T. This may be explained by the fact that τ encodes the characteristic *time* of the relevance decay and therefore cannot be accurately estimated when publications are added within a relatively narrow time window. We, hence, do not report the error estimates of τ in our results presented in Table II.

The finding that, for the growth of the citation network, the parameter estimates are stable over time is not trivial because our analysis of JHEP spans 20 years. Between 1997 and 2017 the scientific landscape has changed considerably. In particular, the pace of academic publishing has been increasing exponentially over the past decades [22]. This accelerated growth in publications (and citations) does not affect our analysis because we have parametrized time in the network growth process by means of new publications. When mapping the growth process to physical time, the earlier time windows in Fig. 6 span a much longer physical time interval compared to later time windows. The fact that we achieve stable parameter estimates in our analysis supports the recent finding that attention decay in science is driven by the accumulation of newer publications rather than by the passing of time [2].

Going back to the results for the nine networks in Table II, we find that model selection for most of the studied journals favors *coupled models* of citation growth, i.e., models accounting for both a *citation* and a *social* constituent. The citation constituent was best described by PA-RD, whereas for the social constituent different candidates were chosen (mostly NPUB).

The three exceptions are the journals Physical Review A (PRA), Physical Review C (PRC), and Physical Review E (PRE), published by the American Physical Society. For these, model selection favors the nonlinear preferential attachment, PA-NL-RD. According to our research design presented in Table I, we excluded tests of PA-NL-RD together with social constituents. Thus, we cannot argue about how PA-NL-RD would fare in coupled models. But we already see that for PRA, for which PA-NL-RD was selected, the exponent $\gamma = 1.14 \pm 0.224$ is not significantly different from 1, i.e., from the linear preferential attachment, PA-RD. Hence, we can suspect that in this case a coupled model does not give an advantage in the model selection; otherwise, it had been selected.

As a second insight, we note that model selection, for *coupled models*, always favored an *additive* coupling between citation and social constituents. The weight of the social constituent measured by $(1-\alpha)$, ranges between 0.07 and 0.19 across networks.

Only for RMP is the selection of additive models not strongly conclusive. It led to the largest relative likelihood $w_m = 0.71$ for the additive coupling, PA-RD+NPUB, but also to a small, yet considerable, $w_m = 0.03$ for its multiplicative counterpart, PA-RDxNPUB.

As a third insight, the social constituent NPUB based on the whole team of authors is always selected over the variant MAXPUB with only the most prominent author. The exception is, again, RMP, where the model PA-RD-MAXPUB accounting for the highest number of publications among authors is selected with relative likelihood $w_m = 0.26$, along with its counterpart that accounts for the number of publications of the whole team PA-RD-NPUB with $w_m = 0.71$.

To summarize our findings, in most of the networks the *coupled growth* model that includes a *social constituent* has a higher likelihood to explain the observed citation dynamics than the simple growth model that only considers the previous history of the citation network.

VII. CONCLUSIONS

In this article, we have provided three contributions: (i) a framework to model the *growth* of networks consisting of *different coupled layers*, (ii) a method for statistical parameter estimation and *model selection* applicable to *growing multilayer networks*, and (iii) a large-scale case study of the citation dynamics in nine different physics journals which demonstrates the applicability of our approach.

To model the growth of multilayer networks is challenging because it combines two different dynamics, *within* and *between* network layers. The coupling between the layers usually does not allow us to separate the timescales for these dynamic processes. Our modular approach provides a convenient way to cope with this in a stochastic manner. Importantly, it allows to encode different *hypotheses* about these dynamics, i.e., to generate different dynamic models, which can then be tested *statistically* against empirical data.

This methodology requires two steps: (i) for each model, the parameters that match the data best have to be obtained using a MLE, and (ii) models with different complexity (i.e., number of free parameters) have to be compared in their ability to describe the data. These two steps bear, again, a number of challenges that we address in this paper. First, the MLE has to be applied to a growing network with coupled dynamic processes. Here we followed a microscopic approach (see also Refs. [19,20]), i.e., we focus on the temporal sequence of edges being added to the network. Second, we need to ensure the temporal stability of the model, i.e., the involved parameters should not change drastically over time. And third, the MLE needs to be computationally efficient, not only for computing parameters but also for computing their standard errors. This requires a scalable procedure that also works for large multilayer networks. We solved this problem by means of sampling of growth events in the network. Eventually, we need to provide an efficient procedure for *model selection*, i.e., for comparing models of different complexity. Here we have chosen relative likelihoods as the most insightful measure (see also Ref. [20]).

In addition to the methodological contributions, our paper also provides interesting insights into scientometrics. Specifically, we addressed the question to what extent social constituents, such as the number of *previous* coauthors or publications, impact the citation of publications. This question was recast in the coupled growth of a two-layer network, where one layer captures the publications and the second one the authors. To construct such dynamic networks, we used data from nine physics journals, where each journal was captured in one two-layer network. We proposed 11 models for the growth of the citation networks, which combine different forms for a citation and a social constituent. These constituents represent different hypotheses about citation growth, mostly from the family of preferential attachment models, and about social influences.

Our first insight regards the important role of the social constituent. We found that in the majority of the studied networks, the data are best described by *coupled* models, i.e., models that incorporate the coupling with the social constituent. Among these, the additive coupling is selected in most of the cases, with the mixture weight of the social component between 7% and 19%. Second, regarding the type of social influence, we learned that the total number of previous coauthors explains the citation data best in most cases. This means that there is no strong statistical evidence for Merton's conjecture [15] that the scientific community tends to cite a publication merely because of the most prominent author with the largest number of previous coauthors. Third, it was interesting to note that not for all of the nine journals the citation dynamics was best described by the *same* model. There was clear evidence that preferential attachment with attention decay is the most promising candidate to explain the observed citations, which was already discussed in the literature [2]. But three journals were better described by nonlinear preferential attachment models, and there was also a variation in the form of the social influences.

Using the modeling framework provided in this paper, our analysis can be extended in different ways. While our focus in this paper was on the growth of citations, the next step could incorporate a model component for authorship formation into the analysis. Such a comprehensive model will facilitate further our understanding about the simultaneous coevolution of citations and authorship relations. It may shed more light on the feedback mechanisms between network layers and lead to insights about successful career paths of authors.

APPENDIX A: MLE AND MODEL SELECTION

In general, we cannot analytically calculate the log-likelihood function defined in Eqs. (2) and (8). The reason is that, e.g., even for the simplest linear preferential attachment, Eq. (19), the normalization factors differ between summands $\pi_{ij}^{\mu\nu}[G(t);\boldsymbol{\theta}]$ for different $\mu\nu$ and t in Eqs. (2) and (8). Hence, we have to resort to numerical computation. However, we can expect that the network growth models studied in this article and in the literature are described by a convex likelihood function with a unique maximum, as discussed in Ref. [20]. We visually confirm this in Fig. 7 for one of our models that implements additive coupling between the citation and social constituents `PA-RD-NCOAUT`.

In Sec. IV, we applied our method for the statistical evaluation of network growth models to synthetic networks with known growth mechanisms. This way we could confirm that the method correctly recovers the parameters of the true generating model. In Table III, we present the results from estimating multiple growth models to such a network. We see that the model selection based on relative likelihoods defined in Eq. (6) correctly identifies the true model among multiple candidates.

APPENDIX B: MODELLING DETAILS FOR ONE EMPIRICAL NETWORK

In Table II we presented the results of model selection for nine empirical networks, for each considering a set of 11 candidate models. Here, in Table IV, we present the detailed outcome of fitting these 11 models to one of the networks, namely JHEP. We report the resulting log-likelihood of the model per edge, i.e., $\ln \mathcal{L}/|E^{\rm pc}|_s$, where $|E^{\rm pc}|_s = 60131$ is the number of edges in the considered sample. This means that each of the publications sampled for model evaluation cites on average $|E^{\rm pc}|_s/5000 \approx 12$ other publications *within* the journal. Recall that even though we use a relatively small sample of publications (about a third in case of JHEP, see

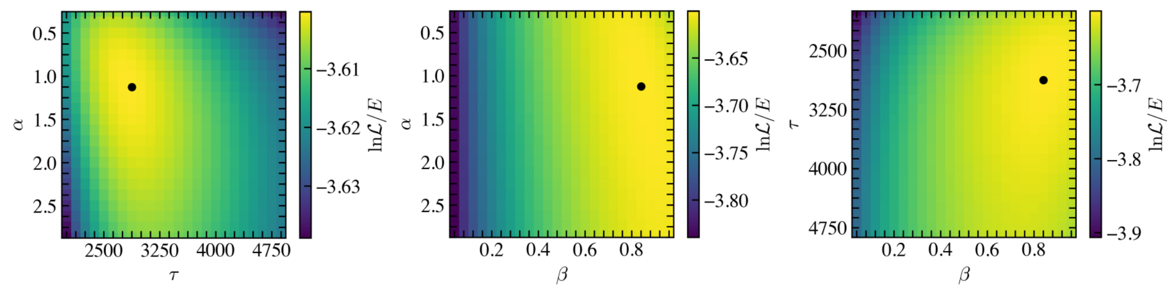

FIG. 7. The log-likelihood function per edge for the model `PA-RD-NCOAUT` fitted to the network of JHEP. Each section plane of the two model parameters shown corresponds to the maximum likelihood value of the third parameter. The dot indicates the location of the maximum likelihood. It is the same in all three plots.

Table VI), *all* their citations to any publication in the whole network are considered.

First, we look at models of citation growth that implement only the citation constituent, i.e., no influences of social constituent capturing the author layer. By comparing the log-likelihoods and the AICs, we find that the two models *with* relevance decay, `PA-RD` and `PA-NL-RD`, considerably outperform the simple preferential attachment model, `PA`, which favors older nodes and, hence, does not reflect their decaying relevance.

The two models `PA-RD` and `PA-NL-RD` estimate a characteristic relevance decay time $\tau \approx 2700$. Recall that we measure time in terms of newly added publications, so the value of τ gives the number of publications that need to be added to the network to reduce the relevance of a given publication by a factor $e \approx 2.72$. From the two models with relevance decay, `PA-RD` describes the data better.

The nonlinear preferential attachment model, `PA-NL-RD`, indicates that $\gamma = 0.93 \pm 0.005 < 1$. This means that the citation history matters less than expected from a simple preferential attachment model. In `PA`, older publications have had more time to accumulate citations and hence tend to have higher in-degree only due to their higher age. To compensate for that, we find from our parameter estimation that for `PA` the constant $\delta = 8.47 \pm 0.188$ has a very high value. Given that δ increases the chance of being cited in particular for new publications, Eq. (17), its high value also mitigates the relevance of older publications.

Next, we see that all models *with* a social constituent perform better than models without it, reflected by the smaller values of both the log-likelihood and the AIC score. The model that best describes the data, with a relative likelihood $w_m = 1.00$, is `PA-RD-NCOAUT`, which accounts for the total number of previous coauthors of the authors of the cited publication. The estimate of the mixture parameter $\alpha = 0.85 \pm 0.007$ between the citation and the social constituents means that approximately 85% of the received citations are driven by previously existing citations and 15% are driven by properties of the authors, namely the counts of their of previous coauthors.

APPENDIX C: DATA

In the following, we describe the empirical data used in our analysis. We used data from two sources, American Physical Society (APS) [31] and INSPIRE [32]. We imported the data entries relevant for our study into a `Neo4j` graph database. Storing the data as a graph with attributes attached to the nodes and edges allowed us to perform the needed constrained graph traversals very efficiently. For instance, for the model `PA-RD-NCOAUT` this data storage format allows to easily count paths of up to length three that traverse a certain sequence of node types (`publication` → `author` → `publication` → `author`) satisfying given constraints: The publications on the path must be in a given journal and the time stamps of subsequent publications on the paths must have earlier timestamps.

1. APS journals

The American Physical Society has been publishing physics journals since 1893. From its foundation to 1970, the *Physical Review* journal published articles on all fields of physics. The format of short letters was introduced in 1958 together with the journal *Physical Review Letters*, aiming to publish notable findings from all areas of physics in a rapid fashion. In 1970 *Physical Review* was split into four journals

TABLE III. Model selection for a synthetic network defined in Sec. IV.

Model	$\ln \mathcal{L}/\lvert E^{pp}\rvert_S$	w_m	α	δ	γ	τ
PA	−2.74568	0.00	—	2.86 ± 0.170	—	—
PA-NAUT	−2.74001	**1.00**	0.72 ± 0.492	1.35 ± 0.385	—	—
PA-RD	−2.74631	0.00	—	2.33 ± 0.123	—	3760 ± 11
PA-NL-RD	−2.75834	0.00	—	—	0.96 ± 0.015	No convergence
PA-RD-NAUT	−2.75159	0.00	0.75 ± 0.080	0.50 ± 0.160	—	No convergence

TABLE IV. Fitting growth models to the network of JHEP based on a sample of 5000 publications that create $|E^{\mathrm{pc}}|_s = 60131$ citations.

Model	$\ln \mathcal{L}/|E^{pp}|_S$	AIC$_m$	w_m	α	β	γ	δ	τ
UNIF	−3.86236	464495	0.00	—	—	—	—	—
PA	−3.75232	451264	0.00	—	—	—	8.47 ± 0.188	—
PA-RD	−3.60539	433595	0.00	—	—	—	2.08 ± 0.057	2735
PA-NL-RD	−3.61012	434164	0.00	—	—	0.93 ± 0.005	—	2644
PA-RD-NAUT	−3.60267	433270	0.00	0.82 ± 1.000	—	—	0.77 ± 1.000	2749
PA-RD-NCOAUT	−3.60015	432968	**1.00**	0.85 ± 0.007	—	—	1.11 ± 0.074	2828
PA-RD-MAXCOAUT	−3.60099	433068	0.00	0.86 ± 0.299	—	—	1.17 ± 2.546	2786
PA-RD-NPUB	−3.60229	433224	0.00	0.90 ± 0.012	—	—	1.42 ± 0.437	2764
PA-RD-MAXPUB	−3.60313	433325	0.00	0.92 ± 0.611	—	—	1.57 ± 1.584	2779
PA-RDxNCOAUT	−3.60167	433150	0.00	—	0.14 ± 0.010	—	2.38 ± 0.471	2794
PA-RDxNPUB	−3.60313	433326	0.00	—	0.10 ± 0.006	—	2.27 ± 0.051	2740

addressing different fields in physics. Later splits and additions of new journals led to the existence of 12 APS journals as of 2018. From the set of historical and currently active journals, we studied the citation and authorship networks of the following five journals: *Physical Review* (PR), *Physical Review A* (PRA), *Physical Review C* (PRC), *Physical Review E* (PRE), and *Reviews of Modern Physics* (RMP). This set covers a wide range of the physics subfields and types of publications. RMP is special in this selection as it focuses on review publications that consolidate the current state of a research topic instead of contributing original research outcomes.

APS freely provides the data on all their publications *"for use in research about networks and the social aspects of science."* For each publication, these data include the title, the names and affiliations of the authors, the relevant time stamps (such as date received, date published), the journal in which it is published, and the list of other APS publications that it cites. The major problem of the raw data for our analysis was that the authors are not uniquely identified, i.e., the same name may describe multiple authors, or the same author may be represented in the data set by different name spelling, e.g., with the first name initial provided with one publication and the fully spelled first name with another publication. To overcome this problem, we augmented the raw dataset with data on author disambiguation obtained using machine learning techniques [33]. Table V summarizes the networks of the five selected journals by the numbers of publications $|V^p|$, authors $|V^a|$, citations among publications $|E^{pp}|$, and the authorship relations $|E^{ap}|$.

2. INSPIRE

This data set contains only publications relevant in the field of high-energy physics, thus allowing us to study how citation behavior differs between journals on a single topic. The underlying INSPIRE online information system is developed by a collaboration of CERN, DESY, Fermilab, IHEP, and SLAC. It consolidates information on high-energy physics publications, researchers, collaborations, and research data and allows access through a web interface, API, and bulk downloads. The major advantage of this data set is the high quality of author disambiguation, which is facilitated by personalized features, such as author profiles and paper claiming.

High-energy physics differs from other fields in one crucial aspect: the significant number of publications with a huge number of authors. As shown in Fig. 8, the INSPIRE data set features 1278 publications with more than a thousand authors. Based on the preceding data set SPIRES, it was pointed out [34] that the authors in high-energy physics have on average 173 coauthors, while in other fields the number ranges between 3.87 and 18.1. Such publications with large numbers of authors relate mostly to collaborations, e.g., the *ATLAS Collaboration*, on large-scale experiments, such as *CERN-LHC-ATLAS*.

Publications with enormously long author lists are *not* suited to study how the citation dynamics is influenced by social constituents, simply because of the dominating number of coauthors. Hence, we needed to focus a subset of the data to exclude such publications. Specifically, from the INSPIRE data set we included publications identified as theoretical or general physics and excluded those identified as experimental, based on corresponding tags present in the original raw data. From the remaining subset of the INSPIRE data, we take the three journals that have the most, specifically more than 10 000, within-journal citations. These are the journals *The Journal of High Energy Physics* (JHEP), *Physics Letters A*

TABLE V. Network summary for five journals published by the APS.

Network	$|V^p|$	$|V^a|$	$|E^{pp}|$	$|E^{ap}|$
Physical Review (PR)	46728	24307	253312	87386
PRA	69147	41428	416639	144806
PRC	36039	22672	253948	108844
PRE	49118	36382	182701	95796
RMP	3006	3788	5282	5044

TABLE VI. Network summary for the four largest journals in the Inspire-HEP data set.

Network	$|V^p|$	$|V^a|$	$|E^{pp}|$	$|E^{ap}|$
JHEP	15739	7994	191990	39056
PR-HEP	44829	33908	213625	115237
Phys. Lett.	22786	18078	56332	53089
Nuc. Phys.	24014	18733	125252	60018

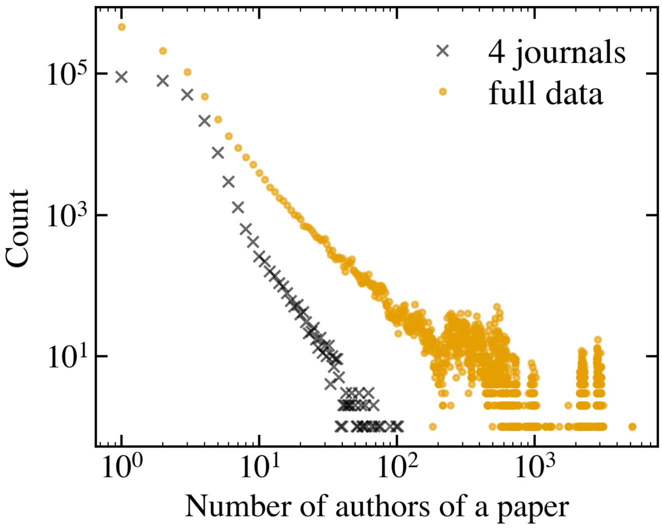

FIG. 8. Distribution of the number of authors of a publication. Yellow dots correspond to the full data set of INSPIRE, black "x" markers correspond to the subset from the three journals JHEP, Phys. Lett., and Nuc. Phys., plus PR-HEP, used in our analysis.

and *B*, and *Nuclear Physics* (Nuc. Phys.), for which we constructed the three networks for our analysis. The distribution of the number of authors for these journals is also shown in Fig. 8.

3. Comparing APS and INSPIRE

The three networks obtained from INSPIRE are to be compared with a fourth network, High Energy Physics in Physical Review Journals (PR-HEP), which includes publications in all APS journals that were relevant enough in high-energy physics to be indexed in INSPIRE. For comparison, we have applied the same selection criteria for PR-HEP as for the subset containing data from JHEP, Physics Letters A, B (Phys. Lett.), and Nuclear Physics (Nuc. Phys.). Hence, studying the network from PR-HEP, which also captures more than 10 000 citations, provides a different perspective on the citation dynamics in the APS journals. Table VI provides summary statistics for the four networks.

[1] M. Wang, G. Yu, and D. Yu, Physica A **388**, 4273 (2009).
[2] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato, J. Informetrics **9**, 734 (2015).
[3] M. Golosovsky and S. Solomon, Phys. Rev. E **95**, 012324 (2017).
[4] M. Wang, G. Yu, and D. Yu, Physica A **387**, 4692 (2008).
[5] M. Golosovsky and S. Solomon, J. Stat. Phys. **151**, 340 (2013).
[6] M. Golosovsky, Phys. Rev. E **97**, 062310 (2018).
[7] D. J. D. S. Price, Science **149**, 510 (1965).
[8] D. D. S. Price, J. Am. Soc. Inf. Sci. **27**, 292 (1976).
[9] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2001).
[10] D. Wang, C. Song, and A.-L. Barabási, Science **342**, 127 (2013).
[11] K. W. Higham, M. Governale, A. B. Jaffe, and U. Zülicke, Phys. Rev. E **99**, 060301(R) (2019).
[12] F. Wu and B. A. Huberman, Proc. Natl. Acad. Sci. USA **104**, 17599 (2007).
[13] M. Medo, G. Cimini, and S. Gualdi, Phys. Rev. Lett. **107**, 238701 (2011).
[14] V. Nanumyan, C. Zingg, and F. Schweitzer, arXiv:1909.13507.
[15] R. K. Merton, Science **159**, 56 (1968).
[16] M. V. Tomasello, G. Vaccario, and F. Schweitzer, EPJ Data Sci. **6**, 22 (2017).
[17] K. Börner, J. T. Maru, and R. L. Goldstone, Proc. Natl. Acad. Sci. USA **101**, 5266 (2004).
[18] E. Sarigol, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, EPJ Data Sci. **3**, 9 (2014).
[19] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)* (ACM, New York, 2008), pp. 462–470.
[20] M. Medo, Phys. Rev. E **89**, 032801 (2014).
[21] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[22] V. Larivière, È. Archambault, and Y. Gingras, J. Am. Soc. Inf. Sci. Tech. **59**, 288 (2008).
[23] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy, Phys. Rev. Lett. **111**, 058701 (2013).
[24] A. Garas (Ed.), *Interconnected Networks* (Springer, Berlin, 2016).
[25] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, SIAM J. Sci. Comput. **16**, 1190 (1995).
[26] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, ACM Trans. Math. Softw. **23**, 550 (1997).
[27] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, Nature Methods **17**, 261 (2020).
[28] T. Pham, P. Sheridan, and H. Shimodaira, Sci. Rep., **6**, 32558 (2016).
[29] P. Sen, Physica A, **346**, 139 (2005).
[30] R. Leimu and J. Koricheva, Trends Ecol. Evol. **20**, 28 (2005).
[31] Data provided by American Physical Society for research purposes, see, https://journals.aps.org/datasets.
[32] See http://inspirehep.net/info/general/project/index for general information and http://inspirehep.net/dumps/inspire-dump.html for data acquisition.
[33] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, Science, **354**, aaf5239 (2016).
[34] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).