


# Quantifying scientists' research ability by taking institutions' scientific impact as priori information

Journal of Information Science  
1–18  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/01655515231191231  
journals.sagepub.com/home/jis  


**Shengzhi Huang** 

School of Information Management, Wuhan University, China; Information Retrieval and Knowledge Mining Laboratory, Wuhan University, China

**Wei Lu** 

School of Information Management, Wuhan University, China; Information Retrieval and Knowledge Mining Laboratory, Wuhan University, China

**Yong Huang**

School of Information Management, Wuhan University, China; Information Retrieval and Knowledge Mining Laboratory, Wuhan University, China

**Zhuoran Luo** 

School of Information Management, Wuhan University, China; Information Retrieval and Knowledge Mining Laboratory, Wuhan University, China

## Abstract

Scholar performance evaluation is extremely important in research assessment decisions, such as funding allocation, academic rankings, and academic promotion. In this article, we propose the institution Q model (IQ) and its two variants (IQ-2 and IQ-3), which aim to evaluate the individual-level research ability to publish high-quality scientific papers. Specifically, our models integrate scientists' institutions, countries and collaborators as valuable prior information and jointly evaluate the research ability of scientists from different institutions. To estimate model parameters and hidden variables defined in our models, we propose a generic BBVI-EM algorithm. To test the effectiveness of our models, we examine their performance on the synthetic data and the empirical data (17,750/26,992 scientists in the computer science/physics field). We find that our models can more accurately quantify the research ability of scientists and institutions and more effectively predict scientists' scientific impact (the h-index and total citations) than the Q model and common machine learning models. In conclusion, our models are effective evaluation and prediction tools for quantifying research ability and predicting the scientific impact, and the BBVI-EM algorithm is an effective variational inference algorithm. This study makes a theoretical contribution to broaden the idea of incorporating the academic environment into scientific evaluation.

## Keywords

Citation analysis; probabilistic graphical model; research ability; variational inference

## 1. Introduction

With the rapid development of science and technology, the number of publications has increased rapidly, and more and more scientists are devoting themselves to scientific research [1]. Thus, studying author-level metrics effectively evaluating scientists' research performance has become an increasingly hot topic because of its practical importance for critical decisions in science, such as funding allocation, academic rankings and academic promotion [2–4].

The existing studies mainly present three approaches for evaluating scientists' research performance. The first type is citation-based metrics [1,5], which require less information and are easy to compute. The second type is network-based metrics [6,7], which analyse the topology structure of the citation network and/or authorship network. The third type is

---

### Corresponding author:

Yong Huang, School of Information Management, Wuhan University, Wuhan 430072, P.R. China.  
Email: yonghuang1991@whu.edu.cn

altmetrics-based metrics [8,9], which use media information from social platform and/or academic platform. However, previous studies still suffer from the following shortcomings: some metrics are criticised for the weak theoretical basis and lack of interpretation, such as the h-index, which simply combine two quantities with unrelated meanings; some metrics generally suffer from the time bias problem, and therefore it is difficult to use them to compare scientists from different ages, such as cumulative citations; and some metrics cannot effectively cope with sparse data, and neglect academic environment [1].

Recently, Sinatra et al. [10] proposed the Q model, in which the  $Q$  parameter can truly account for the individual-level research ability to publish high-quality scientific papers and is not affected by time. However, the Q model neglects scientists' academic environment, such as their institutions, countries and collaborators. Actually, working at a prestigious institution drives greater research performance among early-career researchers [11]. Way et al. [12] further found that the characteristics of a prestigious institution facilitate productivity and prominence by providing a conducive working environment. Previous studies show that collaboration has an effect on scientists' research performance [13–15]. Moreover, the Q model independently evaluates the research ability of scientists and, therefore, cannot effectively cope with sparse data encountered frequently in scientific evaluation. In this study, we propose an explainable generative process to comprehensively consider the academic environment, research ability, and randomness and present a probabilistic graphical model to quantify research ability, which can effectively solve the above shortcomings while retaining the merits of the Q model.

More specifically, we present the novel institution Q model and its two variants, which can integrate scientists' institutions, countries and collaborators as prior information, and jointly evaluates the research ability of scientists from different institutions. To estimate the model parameters and hidden variables defined in our models, we also present the BBVI-EM algorithm. Subsequently, we show that our models achieve better quantification performance of the research ability of scientists and institutions than the Q model on the synthetic data and achieve a better prediction performance of the h-index and total citations than the Q model and common machine learning models on the empirical data (17,750/26,992) scientists in the computer science/physics field extracted from Microsoft Academic Graph data [16].

The current study has the following theoretical and practical implications. We propose the novel institution Q model with an explainable generative process to comprehensively consider the academic environment, research ability and randomness in the citation process. Our models are not only an effective evaluation tools to quantify the research ability of scientists and institutions but also practical prediction tools to predict scientists' scientific impact. This study broadens the idea of how to incorporate the academic environment into scientific evaluation, and other researchers can draw lessons from our modelling methods. Moreover, we propose a universal and effective BBVI-EM algorithm, which can also be used in the inference and estimation of probabilistic graphical models. Finally, at the end of this article, we also offer practical guidelines for our models.

The rest of the article is organised as follows. In Section 2, we review related studies. In Section 3, we introduce the institution Q model and the BBVI-EM algorithm. In Section 4, we clearly introduce the process of generating the synthetic data and the process of collecting the empirical data. In Section 5, we provide an analysis of the experimental results. In Sections 6 and 7, we discuss the contributions and limitations of our study.

## 2. Background

### 2.1. Reviews on author-level evaluation metrics

Scholar performance evaluation is extremely important in research assessment decisions [2–4], and might comprise not only scientists' scientific impact but also their course feedback, educational background, and funding experience [17]. In this article, we focus on reviewing the author-level metrics evaluating scientists' research performance, which can be roughly divided into the following three categories: citation-based methods, network-based methods and altmetrics-based methods.

The publication and citation counts are the most commonly used indicators to evaluate a scientist's productivity and impact [17,18]. The h-index simultaneously gauges scientists' productivity and scientific impact [19,20]. Previous studies have shown that the evaluation results based on the h-index are consistent with the peer review to a certain extent [21]. However, the h-index suffers from some obvious shortcomings. For example, the h-index puts newcomers at a disadvantage, lacks sensitivity to performance changes, and cannot completely capture the distribution of citation frequencies [22]. Therefore, a series of variants of the h-index, such as the g-index [23] and the hg-index [24], have been proposed and analysed. Abbasi et al. [13] employed the g-index as a performance measure and found that scholars with more collaborators have higher g-index by social network analysis. Lungeanu et al. [15] found that scholars with lower h-index collaborate more in interdisciplinary teams. However, the h-index and its variants still have a weak theoretical basis because

they combine two quantities with unrelated meanings. Recently, Zhou et al. [1] presented a simple ranking index which considers the competition among scientists. They demonstrated that the new index achieves better prediction performance in identifying prize-winning scientists than the h-index and the g-index. In conclusion, the raw publication counts and raw citation count-based metrics are generally simple, effective and easy to understand. However, they ignore the citation patterns under different topics and are criticised for their weak theoretical basis.

Network-based methods have also been widely utilised in evaluating scientists, in which scientists as well as papers are generally defined as nodes, while citations, authorship and/or other bibliographic information are defined as edges. For example, Radicchi et al. [25] proposed a PageRank-based algorithm to rank scientists in a citation network. They showed that their score has better predictive power in terms of the assignment of major prizes and awards than total citation counts. Senanayake et al. [26] employed the PageRank algorithm to present a novel p-index based on the citation network, which builds a fairer ranking of scientists compared with the h-index. Recently, Bioglio et al. [27] proposed a novel inspiration score, which quantifies the citation rates of papers. They showed that the inspiration score is an effective index to detect the most inspiring scientists and papers in a citation network. In addition, Liu et al. [7] proposed the AuthorRank index based on the co-authorship network, representing an obvious advantage over degree, closeness and betweenness. Jiang et al. [28] presented hierarchical and non-hierarchical models to quantify the social influences of scientist groups, in which multiple types of collaboration relationships are considered. Overall, network-based methods use direct and indirect link relations and link weights between nodes to disclose the importance of scientists in the network. However, obtaining accurate and complete networks is challenging, and network-based ranking methods often have high computational complexity [2].

Many researchers employ altmetrics to evaluate scientists and papers. Haustein et al. [29] examined the use of academic profile platforms of a sample of bibliometricians, and found that altmetrics indeed reflect impact not reflected in citation counts. Mikki et al. [9] analysed the profiles of 4,307 scientists affiliated to the University of Bergen, and showed that the correlation between the traditional bibliometric indicators and social activity indicators is low. Ortega [30] analysed about 10,000 scholar profiles from the Spanish National Research Council, and found that there is little correlation between altmetrics and literature metrics at the author level. Martín-Martín et al. [8] examined the profile of 811 scientists from the bibliometrics field, and analysed 31 author-level metrics collected from academic profile platforms. They found that altmetrics and citation metrics have different focuses. Fang et al. [31] investigated how frequently short links to scientific papers are clicked on Twitter, and found that Twitter clicks are weakly correlated with scholarly impact metrics. Lemke et al. [32] recently found that many researchers exhibit a certain scepticism about altmetrics. Taken overall, altmetrics tend to measure the social impact and social ability, which is not the focus of this study.

Recently, Sinatra et al. [10] proposed the Q model, in which the  $Q$  parameter can truly account for a scientist's research ability. However, their model neglects the academic environment of scientists, and cannot cope with sparse data encountered frequently in scientific evaluation, which confines its prediction power. In this article, we provide a series of better models for scientific research evaluation.

## 2.2. Reviews on probabilistic graphical model and its inference algorithm

Probabilistic graphical models with latent variables have been widely used in modern learning applications, such as text mining, image processing, and information retrieval [33–38]. Probabilistic graphical models postulate a meaningful generative process responsible for the observation ( $X$ ), infer the hidden variables ( $Z$ ), draw conclusion from the observed data, and make predictions about new data [39]. The key issue of probabilistic graphical models is to derive the posterior distribution of the hidden variables ( $P(Z|X)$ ), which is generally intractable [40]. Therefore, practitioners generally resort to two common approximate methods (i.e. Markov Chain Monte Carlo and variational inference) for estimating the posterior distribution.

The basic idea of Markov Chain Monte Carlo is to sample a set of samples to approximate the target distribution, in which a Markov chain is constructed to converge to the target distribution, and samples are sampled from that Markov chain [41]. For example, Griffiths and Steyvers [42] computed full conditional distribution of the Latent Dirichlet allocation model (LDA),  $P(Z_i|Z_{-i}, X)$ , and employed the Gibbs sampling algorithm to sample from  $P(Z_i|Z_{-i}, X)$ . They utilised the samples to estimate the parameters in LDA. Rosen-Zvi et al. [43] presented the Author-Topic model, in which authorship information is included. They also utilised Gibbs sampling to estimate the topic and author distribution. However, to ensure the Markov chain converges, the chain should be run for enough iterations, which is time-consuming.

The basic idea of variational inference is to use the Jensen inequality to obtain the Evidence Lower Bound (ELBO) of the log likelihood ( $\text{ELBO} \leq \log P(X, Z)$ ), and find the member of a family of variational distributions that is closest to the ELBO in KL distance [44]. However, for generic probabilistic graphical models and arbitrary variational

distributions, there is no closed-form ELBO, which forces practitioners to design model-specific algorithms. The tedious work of designing model-specific algorithms hinders practitioners from rapidly exploring diversified models. Recently, Ranganath et al. [39] proposed a novel ‘black box’ variational inference algorithm (BBVI), which can be easily deployed on any probabilistic graphical models. The core idea of the BBVI algorithm is to employ Markov Chain Monte Carlo to approximate the gradient of the ELBO by sampling from the variational distribution, and use stochastic optimisation to maximise the ELBO. However, the estimated gradient is a little bit problematic, which leads to unstable estimation results. Moreover, Zhao et al. [45] used a variational auto-encoder framework (VAE), which is essentially a neural network, to approximate the posterior of their topic model. Ning et al. [46] also employed the VAE to solve their nonparametric topic model. In this article, we present the BBVI-EM algorithm based on the BBVI algorithm [39] and variational EM algorithm [33], which can effectively solve our models.

### 3. Methodology

#### 3.1. Problem definition

Inspired by previous studies [10,11,47,48], we understand research ability as the inherent ability of scientists to publish high-quality papers by taking advantage of the available knowledge [10]. Unlike the number of publications and citations that can be directly observed and time-varying, a scientist’s research ability is an underlying characteristic, which remains relatively stable over his or her career path and potentially affects productivity and the impact [11].

Quantifying research ability is a direct approach to evaluate scientists and a critical foundation for understanding their research performance. This study aims to propose an effective and explainable evaluation model based on the probabilistic graphical model to quantify scientists’ research ability in which academic environment (i.e. institutions, countries and collaboration) are employed as valuable prior information.

#### 3.2. Review of the Q model

Before introducing our model, first, we simply review the Q model [10]. In the Q model,  $C_{\alpha,i}^{10}$  (the number of citations of a paper  $i$  authored by a scientist  $\alpha$  10 years after publication) is employed to gauge the quality of the paper, and is assumed to be determined by the multiplicative processes between  $Q_\alpha$  and  $p_i$ , in which  $Q_\alpha$  captures the research ability of  $\alpha$ , and  $p_i$  indicates the luck, as shown in equation (1). Hence, a high-impact publication (large  $C_{\alpha,i}^{10}$ ) is published by a scientist with excellent research ability (large  $Q_\alpha$ ) and good luck (large  $p_i$ ). Subsequently, they consider the log-normal nature of  $P(C_{\alpha,i}^{10})$ , denote  $\hat{p} = \log(p)$  and  $\hat{Q} = \log(Q)$ , and obtain the joint probability of ability, luck and productivity, the trivariate normal distribution  $P(\hat{p}, \hat{q}, \hat{N}) \sim N(\mu, \Sigma)$ . Finally, a classical maximum likelihood estimation method is employed to estimate  $Q_\alpha$ , as shown in equation (2).  $\langle \log C_{\alpha,i}^{10} \rangle$  indicates the average value of  $\log C_{\alpha,i}^{10}$  of all papers published by  $\alpha$ . Their results show that  $\hat{p}$ ,  $\hat{Q}$  and  $\hat{N}$  are almost independent of each other, which effectively untangles the role of productivity, luck and ability in a scientific career

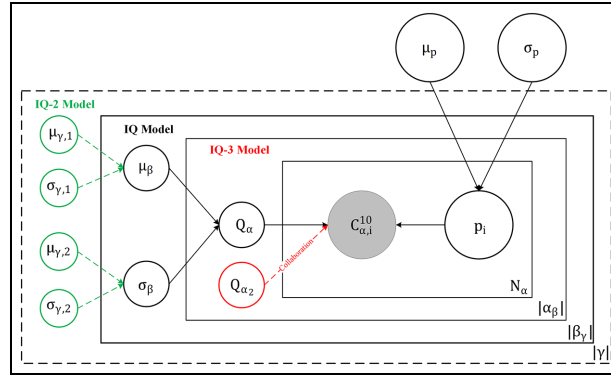
$$C_{\alpha,i}^{10} = Q_\alpha p_i \quad (1)$$

$$Q_\alpha = e^{\langle \log C_{\alpha,i}^{10} \rangle - \mu_p} \quad (2)$$

However, the Q model ignores the academic environment, which may directly or indirectly affect scientists’ research performance and in turn affect the process of quantifying research ability. Actually, scientists from the same institution tend to cooperate frequently, and working at a prestigious institution drives better research performance among scientists by providing a conducive working environment [11,12]. Hence, a scientist’s affiliated institution, the country where the institution is located, and collaboration among scientists are important information for quantifying scientist’s research ability. In addition, the Q model cannot effectively cope with sparse data encountered frequently in scientific evaluation. Actually, young scientists lack enough time to accumulate sufficient publication records and citation records [1]. The goal of this study is to provide a better evaluation model which can effectively solve the above shortcomings.

#### 3.3. The institution Q model

This study proposes the institution Q model (IQ model), which integrates scientists’ affiliated institutions as valuable prior information, and jointly evaluates all scientists from different institutions. The core idea of the IQ model is to assume that the research ability of scientists from the same institution shares the same distribution, by which our model can cope with data sparse in author-level evaluation and also explain the research ability of institutions. This simple but



**Figure 1.** Graphical model representation of the institution Q model.

intuitive idea not surprisingly improves the performance of our model and broadens the thought for incorporating the academic environment into scientific evaluation. We continue to use the notations used in the Q model and define the following explainable generative process for a scientist’s publication sequence and citation records.

The IQ model assumes that there are several institutions (i.e.  $\beta \in \beta$ ), in which there exist many scholars in each institution (i.e.  $\alpha \in \alpha_\beta$ ). For each scientist, individual publication and citation records are generated in the following three steps:

- *Step 1.* For each scientist  $\alpha$  from an institution  $\beta$ , we sample  $\alpha$ ’s productivity from  $\beta$ ’s productivity distribution,  $N_\alpha \sim \text{Poisson}(\lambda_\beta)$ .
- *Step 2.* For the scientist  $\alpha$ , we sample  $\alpha$ ’s research ability from  $\beta$ ’s research ability distribution,  $Q_\alpha \sim \text{LogNormal}(\mu_\beta, \sigma_\beta^2)$ .
- *Step 3.* For each paper  $i$  authored by  $\alpha$ , we sample the luck  $p_i \sim \text{LogNormal}(\mu_p, \sigma_p^2)$ , and use the equation,  $C_{\alpha,i}^{10} = Q_\alpha p_i$ , to generate citation records.

In the above generative process, first, we sample the number of publications of  $\alpha$  ( $N_\alpha$ ) from the Poisson distribution.  $\lambda_\beta$  represents the average productivity of  $\beta$ . Subsequently, we sample  $\alpha$ ’s research ability ( $Q_\alpha$ ) from the LogNormal distribution, in which  $\mu_\beta$  indicates the average research ability of  $\beta$ . Finally, we repeatedly sample  $p_i$   $N_\alpha$  times, and generate  $\alpha$ ’s citation records,  $(C_{\alpha,1}^{10}, C_{\alpha,2}^{10}, \dots, C_{\alpha,N_\alpha}^{10})$ .  $\mu_p$  and  $\sigma_p$  describe the randomness of the citation process in a specific field. Unlike the Q model, the productivity of a scientist also plays a non-ignorable role in the institution Q model; that is, more high-quality articles authored by the scientist will help him or her score higher research ability by twisting known prior information about an institution. The graphical model representation of the institution Q model is shown in Figure 1.

Moreover, we also propose two simple variants of the IQ model. In the first variant (IQ-2 model), we employ the country in which an institution is located as priori information about the research ability of the institution by enriching *Step 2*. Specifically, we assume that the research ability of institutions  $(\mu_\beta, \sigma_\beta)$  from the same country shares the same distribution (i.e.  $\mu_\beta \sim \text{Normal}(\mu_{\gamma,1}, \sigma_{\gamma,1}^2)$  and  $\sigma_\beta \sim \text{LogNormal}(\mu_{\gamma,2}, \sigma_{\gamma,2}^2)$ ). Hence,  $\mu_\beta$  and  $\sigma_\beta$  also become latent variables, as shown by the green circle in Figure 1. In the second variant (IQ-3 model), we consider the collaboration among scientists by modifying *Step 3*. Specifically, we model the quality of a paper as an outcome of joint efforts of Li et al. [11], and use a linear combination of  $Q_\alpha$  of authors to determine  $C_{\alpha,i}^{10}$  (i.e.,  $\log C_{\alpha,i}^{10} = \sum_{i \in \mathcal{A}_i} w_{\alpha,i} \log Q_\alpha + \log p_i$ ), which means that all authors contributed equally to a paper (i.e. fractional count [49]), as shown by the red line in Figure 1.  $w_{\alpha,i}$  indicates the contribution ratio of  $\alpha$  to  $i$  ( $\sum_{i \in \mathcal{A}_i} w_{\alpha,i} = 1$ ), and  $\mathcal{A}_i$  is the set of authors of  $i$ .

In the above models (IQ, IQ-2, and IQ-3),  $N_\alpha$  is independent from  $Q_\alpha$  and  $p_i$  in the generative process, which is based upon the conclusions from the Q model. Hence,  $N_\alpha$  can be seen as an ancillary variable, and we can generally ignore its randomness in the subsequent development [33]. In the IQ model, the joint distribution of the observed data  $(N_\alpha, C_{\alpha,i}^{10}, \beta_\alpha)$ , hidden variables  $(Q_\alpha, p_i)$ , and model parameters  $(\mu_\beta, \sigma_\beta, \mu_p, \sigma_p)$  is shown in equation (3). In the IQ-2 model, the joint distribution of the observed data  $(N_\alpha, C_{\alpha,i}^{10}, \beta_\alpha)$ , hidden variables  $(\mu_\beta, \sigma_\beta, Q_\alpha, p_i)$ , and model parameters  $(\mu_{\gamma,1}, \sigma_{\gamma,1}, \mu_{\gamma,2}, \sigma_{\gamma,2}, \mu_p, \sigma_p)$  is shown in equation (4). In the IQ-3 model, the observed data, hidden variables and model parameters are the same as those in the IQ model, but *Step 3* is different, as shown in equation (5).

**Table 1.** The pseudo code of the BBVI-EM algorithm.

Input:	Data ( $X$ ); Model parameters ( $\psi_0$ ); Variational parameters ( $\phi_0$ ); Number of iterations
Output:	Model parameters ( $\psi$ ); Variational parameters ( $\phi$ )
1	$iters = 0;$
2	<b>While True</b>
3	$iters += 1; E\_iters = 0; M\_iters = 0;$
4	<b>While True</b> # update variational parameters (E-step)
5	$E\_iters += 1$
6	$\phi_m = \phi_{m-1} + \rho \nabla \mathcal{L}(\phi_{m-1})$
7	<b>If</b> $ \psi_m - \psi_{m-1} _{ED} < \epsilon_1$ <b>or</b> $E\_iters < num\_E\_iters$ <b>then</b> <b>break</b>
8	<b>While True</b> # update model parameters (M-step)
9	$M\_iters += 1$
10	$\psi_m = \psi_{m-1} + \rho \nabla \mathcal{L}(\psi)$
11	<b>If</b> $ \phi_m - \phi_{m-1} _{ED} < \epsilon_2$ <b>or</b> $M\_iters < num\_M\_iters$ <b>then</b> <b>break</b>
12	<b>If</b> $( \psi_m - \psi_{m-1} _{ED} < \epsilon_1$ <b>and</b> $ \phi_m - \phi_{m-1} _{ED} < \epsilon_2)$ <b>or</b> $iters < num\_iters$ <b>then</b> <b>break</b>
13	<b>End</b>

BBVI: 'black box' variational inference. EM: Expectation Maximization.

Different from the Q model, the inferential problem of our models is to compute the posterior distribution of the hidden variable ( $Z$ ) given the observed data ( $X$ ),  $P(Z|X)$ . Unfortunately, this distribution is intractable to compute. Hence, we propose the BBVI-EM algorithm to approximate it

$$P(X, Z) = \prod_{\beta} \prod_{\alpha}^{\alpha_{\beta}} P(N_{\alpha}; \lambda_{\beta}) P(Q_{\alpha}; \mu_{\beta}, \sigma_{\beta}) \prod_{i=1}^{N_{\alpha}} P(\log(C_{\alpha,i}^{10}) - \log(Q_{\alpha}); \mu_p, \sigma_p) \quad (3)$$

$$P(X, Z) = \prod_{\gamma} \prod_{\beta}^{\beta_{\gamma}} P(\mu_{\beta}; \mu_{\gamma,1}, \sigma_{\gamma,1}) P(\sigma_{\beta}; \mu_{\gamma,2}, \sigma_{\gamma,2}) \times \prod_{\alpha}^{\alpha_{\beta}} P(N_{\alpha}; \lambda_{\beta}) P(Q_{\alpha}; \mu_{\beta}, \sigma_{\beta}) \prod_{i=1}^{N_{\alpha}} P(\log(C_{\alpha,i}^{10}) - \log(Q_{\alpha}); \mu_p, \sigma_p) \quad (4)$$

$$P(X, Z) = \prod_{\beta} \prod_{\alpha}^{\alpha_{\beta}} P(N_{\alpha}; \lambda_{\beta}) P(Q_{\alpha}; \mu_{\beta}, \sigma_{\beta}) \times \prod_{i=1}^{N_{\alpha}} P\left(\log(C_{\alpha,i}^{10}) - \sum_{i \in \mathcal{A}_i} w_{\alpha,i} \log Q_{\alpha}; \mu_p, \sigma_p\right) \quad (5)$$

### 3.4. Inference and parameter estimation

To approximate the posterior distribution ( $P(Z|X)$ ), we propose the BBVI-EM algorithm based on the variational EM algorithm [33] and BBVI [39]. The BBVI-EM is a generic algorithm that can be easily deployed in our models to optimise iteratively variational parameters ( $\phi$ ) and model parameters ( $\psi$ ), by which  $Q_{\alpha}$  can be estimated. The BBVI-EM alleviates the problematic gradient of ELBO in the BBVI and produces stable estimation results, as shown in Supplemental Appendix 1. We present the pseudo-code of the BBVI-EM algorithm in Table 1.

Specifically, first, we employ the BBVI algorithm to approximate  $P(Z|X)$ . To this aim, a series of variational distribution  $q(Z; \phi)$  with the free variational parameters ( $\phi$ ) are introduced. Our goal is to adjust the variational parameters so that  $q(Z; \phi)$  is close to  $P(Z|X, \psi)$ ; that is to minimise the KL distance between them,  $KL(q(Z|\phi) || P(Z|X, \psi))$ . Notably, minimising the KL distance is equal to maximising the ELBO,  $\mathcal{L}(\phi)$  [41], as shown in equation (6). Taking the IQ model as an example, the specific mathematical form of ELBO is shown in equation (7). For each scientist, a variational distribution,  $qQ_{\alpha} \sim \text{LogNormal}(\mu_{\alpha}, \sigma_{\alpha}^2)$ , is introduced, and therefore there are  $2 * \sum_{\beta} |\alpha_{\beta}|$  variational parameters

$$\mathcal{L}(\phi) = E_{q(z)}[\log P(X, Z) - \log q(Z)] \quad (6)$$

$$\mathcal{L}(\mu_{\alpha}, \sigma_{\alpha}) = E_{q(Q_{\alpha}|\mu_{\alpha}, \sigma_{\alpha})}[\log P(N_{\alpha}, C_{\alpha,i}^{10}, \beta_{\alpha}, Q_{\alpha}) - \log q(Q_{\alpha}|\mu_{\alpha}, \sigma_{\alpha})]$$

**Table 2.** The experimental setup in the synthetic data generation.

Parameter	Simulation 1	Simulation 2	Simulation 3
$ \beta $	14	14	28
$ \alpha_\beta $	10	20	40
$\lambda_\beta$	10	15	20
$\mu_p$	0	0	0
$\log(\sigma_p)$	1	2	3

$$\alpha \in \alpha_\beta; \beta \in \beta \quad (7)$$

Subsequently, the noisy unbiased gradient of  $\mathcal{L}(\mu_\alpha, \sigma_\alpha)$  with Monte Carlo samples from  $qQ_\alpha$  is calculated, as shown in equation (8). A stochastic optimisation is used to maximise the ELBO.  $S$  indicates the number of samples

$$\begin{aligned} \nabla \mathcal{L}(\mu_\alpha, \sigma_\alpha) &= \frac{1}{S} \sum_{s=1}^S \nabla \log q(Z_s) (\log P(X, Z_s) - \log q(Z_s)) \\ Z_s &\sim q(Z; \mu_\alpha, \sigma_\alpha) \end{aligned} \quad (8)$$

Second, we employ the variational EM algorithm to approximate  $P(Z|X)$  by updating the variational parameters and the model parameters iteratively, as shown in lines 6 and 10 of Table 1. The two steps are repeated until the ELBO converges. Finally,  $\mu_\alpha$  can be used to estimate  $Q_\alpha$ , and  $\sigma_\alpha$  can be used to evaluate the accuracy of the estimation.  $\mu_\beta$  explains the research ability of an institution, and  $\sigma_\beta$  reflects the different degree of individual ability of scientists from the institution. Our code is available online.<sup>1</sup>

## 4. Data

### 4.1. Synthetic data generation

One of the important uses of the institution Q model is to evaluate the research ability of a scientist ( $Q_\alpha$ ), which is actually an unknown hidden variable, and its real value can never be known to us in a real scenario. However, in the synthetic data generated by the generative process mentioned above, we can obtain not only the value of observation ( $N_\alpha, C_{\alpha,i}^{10}$ ) but also true value of the hidden variables and model parameters. Hence, first, we evaluate the estimation accuracy of  $Q_\alpha$  on the synthetic data.  $Q_\alpha$  in the Q model is estimated by the maximum likelihood estimation method, and  $Q_\alpha$  in our model is estimated by the BBVI-EM algorithm.

Here, we clarify the experimental setup in synthetic data generation. Specifically, we sample the model parameters of  $|\beta|$  institutions,  $\mu_\beta, \log(\sigma_\beta)$ , from Normal(0, 1). For each institution  $\beta \in \beta$ , we sample the research ability of  $|\alpha_\beta|$  scientists from LogNormal( $\mu_\beta, \sigma_\beta^2$ ). For each scientist  $\alpha \in \alpha_\beta$ , we sample  $N_\alpha$  from Poisson( $\lambda_\beta$ ), and then sample  $p_i N_\alpha$  times from LogNormal( $\mu_p, \sigma_p^2$ ). For simplicity, for each institution,  $\lambda_\beta$  and  $|\alpha_\beta|$  are, respectively, the same. Notably, for each scientist, if  $Q_\alpha$  is increased or decreased by  $\mu_p$ , its relative value is not affected, and therefore  $\mu_p$  is always set to 0. We repeat our experiment at different values of  $|\beta|, |\alpha_\beta|, \lambda_\beta, \sigma_p$ , as shown in Table 2. The simple synthetic data (Simulation 1–3) only supports the information required by the Q and IQ models, by which we compare the quantification performance of the two models. Comparison results on more complicated synthetic data supporting the IQ-2 and IQ-3 models can be found in our Supplemental Appendix 1.

### 4.2. Empirical data collection and preprocessing

Another important use of the institution Q model is the predictive power of  $Q_\alpha$ . To compare the predictive power of the Q model and the IQ, IQ-2 and IQ-3 models on the empirical data, we employ Microsoft Academic Graph (MAG) data [16] as our dataset. FoSs (field of study) generated in the MAG are employed to identify topics of papers. Subsequently, we, respectively, select scientists in the computer science field (CS) and scientists in the physics field to create two empirical datasets according to the following conditions:

1. First, we choose scientists in the CS/physics field, who published their first paper between 1990 and 2000, and have published at least 30 papers until 2010, as research objects. If FoSs of a paper contains ‘*Computer Science*’ (‘*Physics*’), the paper belongs to the CS (physics) field.
2. Second, for simplicity, we identify an affiliated institution of a scientist by counting the number of papers s/he published in her or his institution. This simplification is not necessary for our model. We determine the country where an institution is located through its latitude and longitude. Subsequently, we select institutions with at least 30 scientists and their scientists as our subjects. There are a total of 17,750 (26,992) scientists left in the CS (physics).
3. Third, we need to ensure that each paper has a citation record of at least 10 years; the publication records and citation records of these scientists in the CS (physics) field from 1990 to 2010 are collected.

Notably, to ensure that the prediction experiments resemble the actual prediction situation, we do not employ  $N_1$  and  $N_2$  to split the training data and test data, as Sinatra et al. [10] did, but use the time  $Y_1$  and  $Y_2$  instead. Because each scientist may publish his or her  $N_1$ th paper at different times, Therefore, we use publication records from 1990 to  $Y_1$  as the training data ( $1990 \leq Y_1$ ), and use publication records from  $Y_1$  to  $Y_2$  as the test data ( $Y_2 \leq 2010$ ). Specifically, we use the  $C_{\alpha,i}^{10}$  of papers from 1990 to  $Y_1$  to estimate  $Q_\alpha$ , and use the estimated  $Q_\alpha$  to predict the h-index and  $C_{tot}$  in  $Y_2$ . This can be achieved by sampling  $p_i$  from  $\text{Normal}(\mu_p, \sigma_p)$  for a paper, and calculating  $C_{\alpha,i}^{10}$  by equation (1) based on estimated productivity. Considering that the majority of papers acquire most citations within 2 or 3 years after publication [50], the h-index is approximately computed by  $C_{\alpha,i}^{10}$ , and  $C_{tot}$  is equal to the sum of  $C_{\alpha,i}^{10}$  of all papers authored by  $\alpha$  [10].

## 5. Experiments and results

### 5.1. Evaluation metrics

To evaluate the quantification performance of the research ability ( $Q_\alpha$  and  $\mu_\beta$ ) and the prediction performance of the future impact (the h-index and  $C_{tot}$ ), we employ four popular criteria (*Pearsonr*,  $R^2$ , RMSE, and MAE), as shown in equations (9)–(11). The Pearson correlation coefficient (*Pearsonr*) measures the linear correlation between the predicted value  $\hat{y}_n$  and the real value  $y_n$ . The goodness of fit ( $R^2$ ) gauges the overall relationship between  $\hat{y}_n$  and  $y_n$ . The root mean square error (RMSE) and mean absolute error (MAE), respectively, measure the variation of  $\hat{y}_n$  to  $y_n$  and the average of absolute errors between  $\hat{y}_n$  and  $y_n$ .  $N$  denotes the number of samples.  $\bar{y}_n$  and  $\bar{\hat{y}}_n$  indicate the average of  $y_n$  and the average of  $\hat{y}_n$ , respectively. In the experiments on the synthetic data and the empirical data,  $y_n$  represents  $Q_\alpha$  or  $\mu_\beta$  and the h-index or  $C_{tot}$ , respectively

$$Pearsonr = \frac{\sum_{n=1}^N (y_n - \bar{y}_n)(\hat{y}_n - \bar{\hat{y}}_n)}{\sqrt{\sum_{n=1}^N (y_n - \bar{y}_n)^2 (\hat{y}_n - \bar{\hat{y}}_n)^2}} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y}_n)^2} \quad (10)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2} \quad (11)$$

$$MAE = \frac{1}{N} \sum_{n=1}^N |y_n - \hat{y}_n| \quad (12)$$

### 5.2. Quantification performance analysis

We compare the quantification performance of the research ability ( $Q_\alpha$  and  $\mu_\beta$ ) of the IQ model and that of the Q model on synthetic data. Due to the fact that hidden research ability cannot be observed, common machine learning models are generally unable to cope effectively with this issue for lacking of training data. Therefore, we only compare with the Q model. However, in the Q model,  $\mu_\beta$  is not defined, and we simply employ the average of estimated  $Q_\alpha$  of all scientists from  $\beta$  to estimate  $\mu_\beta$  (i.e.  $\sum_{\alpha \in \alpha_\beta} Q_\alpha / |\alpha_\beta|$ ).

Specifically, we repeat the experiments 20 times under each configuration (*Simulation 1–3*, as shown in Table 2). The t-test is employed to test the significant difference between the estimation accuracy of  $Q_\alpha$  and  $\mu_\beta$  in terms of *Pearsonr*,  $R^2$ , RMSE, and MAE. We report the average of each evaluation metric in Table 3. The ‘bold text’ indicates the best



**Table 3.** Quantification performance on the synthetic data.

Simulation	Metrics	The Q model		The institution Q model	
		$\mu_\beta$	$Q_\alpha$	$\mu_\beta$	$Q_\alpha$
Simulation 1	<i>Pearsonr</i>	<b>0.7865</b>	0.9093	0.7846	<b>0.9330</b> ***
	$R^2$	<b>0.4453</b>	0.7791	0.4451	<b>0.8673</b> ***
	RMSE	0.7317	0.9211	<b>0.7290</b>	<b>0.7282</b> ***
	MAE	0.5506	0.7258	<b>0.4968</b>	<b>0.5563</b> ***
Simulation 2	<i>Pearsonr</i>	0.8341	0.7557	<b>0.8403</b>	<b>0.8586</b> ***
	$R^2$	0.5028	0.2908	<b>0.5124</b>	<b>0.7264</b> ***
	RMSE	0.6585	2.0001	<b>0.6550</b>	<b>1.1652</b> ***
	MAE	0.5065	1.5792	<b>0.5073</b>	<b>0.8764</b> ***
Simulation 3	<i>Pearsonr</i>	0.8090	0.4815	<b>0.8120</b>	<b>0.6738</b> ***
	$R^2$	0.4064	0.0000	<b>0.4216</b>	<b>0.3811</b> ***
	RMSE	0.7559	4.6290	<b>0.7446</b>	<b>1.9252</b> ***
	MAE	0.5996	3.6638	<b>0.5940</b>	<b>1.4077</b> ***

RMSE: root mean square error; MAE: mean absolute error.

\*\*\*p value < 0.001.

result, and ‘\*\*\*’ represents the significant level. Notably,  $R^2$  may take a negative value, because the model can be arbitrarily worse. Therefore, we define  $R^2 = \max(R^2, 0)$ , before averaging  $R^2$ .

As shown in Table 3, in *simulation 1–3*, our model has significantly improved the estimation accuracy of  $Q_\alpha$  in terms of all evaluation metrics. Specifically, in *simulation 1*, *Pearsonr* and  $R^2$  of our model on  $Q_\alpha$  are higher than 2.61% and 11.32% for the Q model. RMSE and MAE of our model on  $Q_\alpha$  are lower than 20.78% and 23.35% for the Q model. In *simulation 2*, *Pearsonr* and  $R^2$  of our model on  $Q_\alpha$  are higher than 13.61% and 149.79% for the Q model. RMSE and MAE of our model on  $Q_\alpha$  are lower than 41.74% and 44.50% for the Q model. In *simulation 3*, *Pearsonr* of our model on  $Q_\alpha$  is higher than 39.94% for the Q model.  $R^2$  of the Q model on  $Q_\alpha$  is always the negative value, and that of our model is the positive value (0.3811). RMSE and MAE of our model on  $Q_\alpha$  are lower than 58.41% and 61.58% for the Q model. Therefore, compared with the Q model, our model can more accurately quantify a scientist’s research ability. In addition, we also find that our model slightly improves estimation accuracy of  $\mu_\beta$  than the average estimation method mentioned above. The average productivity of scientists ( $\lambda_\beta$ ) is low (10, 15, and 20), which means that our model can better cope with sparse data. Moreover, the larger value of  $\log(\sigma_p)$  brings more noise to the observation data, which makes the estimation accuracy of our model on *simulation 3* the lowest and that on *simulation 1* the best.

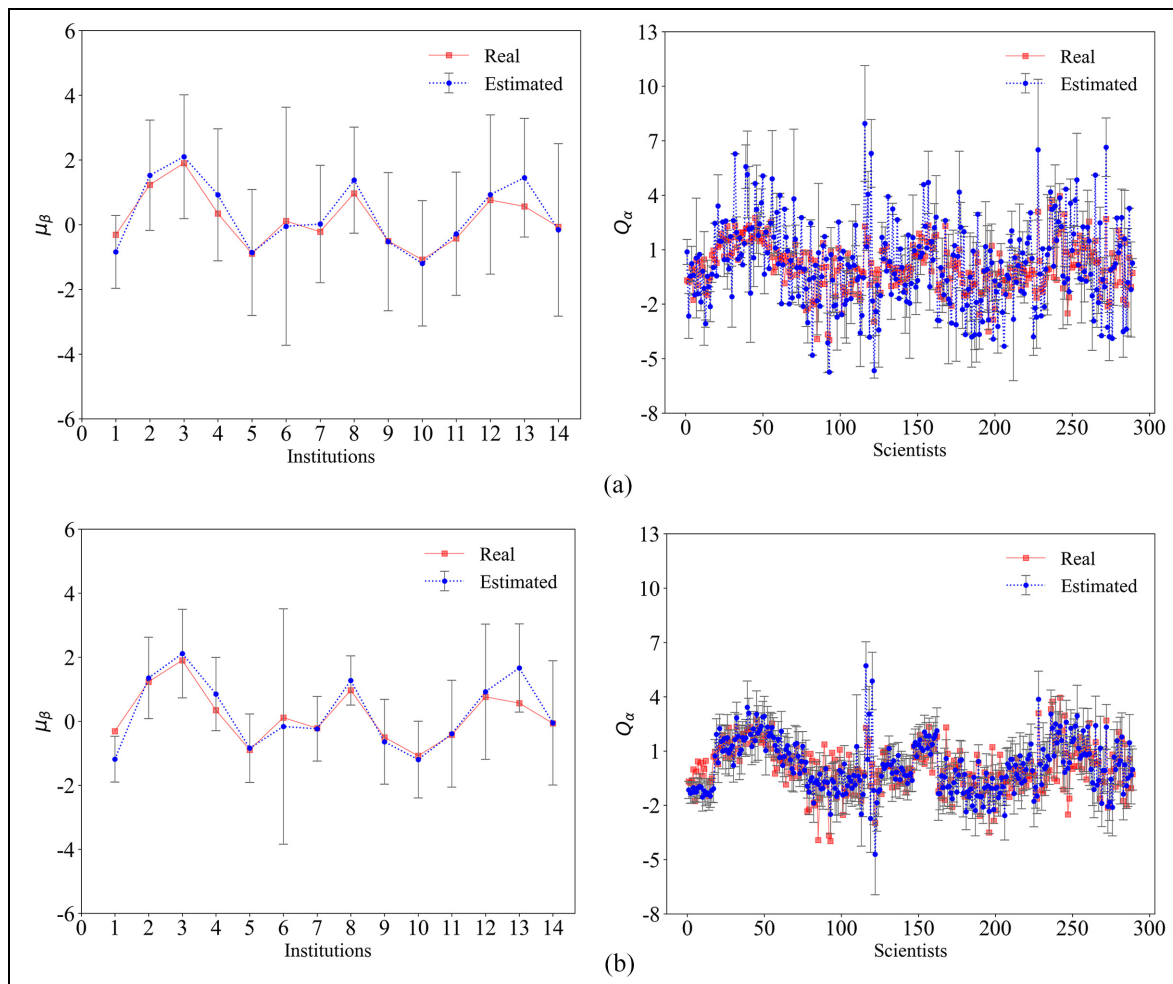
To clearly show the better quantification performance of the IQ model, we randomly select a result from *simulation 2*, and plot the real  $Q_\alpha$  ( $\mu_\beta$ ) distribution and the estimated  $Q_\alpha$  ( $\mu_\beta$ ) distribution. As shown in Figure 2, the red dot indicates the real value for all scientists (institutions), and the blue dot indicates the estimated value. The black error bar denotes the 95% confidence interval of estimated value, and the shorter error bar containing the real value means the more accurate estimation. To be specific, we find that  $Q_\alpha$  estimated by the IQ model is closer to the real  $Q_\alpha$  than those estimated by the Q model. In addition, the error bar in our model is shorter than that in the Q model. For estimated  $\mu_\beta$ , we obtain similar results. We report comprehensive comparative results of Q, IQ, IQ-2 and IQ-3 models on more complicated synthetic data in Supplemental Appendix 1. In conclusion, our model achieves a satisfied quantification performance of the research ability of scientists and institutions, and effectively cope with sparse data in scientific evaluation.

### 5.3. Prediction performance analysis

In this section, we compare the predictive power of our models with the Q model and common machine learning models on the empirical data. The publication records of 17,750 (26,992) scientists in the CS (physics) field is used to carry out the following prediction experiments.

### 5.4. The prediction performance compared with the Q model

We compare the predictive power with the Q model. We use publication records from 1990 to 2000 ( $Y_1$ ) and  $C_{\alpha,i}^{10}$  of those papers as the training data, by which we can estimate  $Q_\alpha$  of scientists. Subsequently, we use the estimated  $Q_\alpha$  to predict the h-index and  $C_{\text{tot}}$  of  $\alpha$  at  $Y_2$ , where  $Y_2 \in [2001, 2010]$ . Specifically, for each scientist, we predict the h-index



**Figure 2.** One case study from simulation 2. (a)  $Q_\alpha$  and  $\mu_\beta$  distributions in the Q model. (b)  $Q_\alpha$  and  $\mu_\beta$  distributions in the institution Q model.

and  $C_{tot}$  at  $Y_2$ , by sampling 100 times and use the average value as the final forecast value. We report our prediction experiments under different  $Y_2$ .

As shown in Tables 4 and 5, we report the prediction accuracy on the h-index and  $C_{tot}$  in terms of *Pearsonr*,  $R^2$ , RMSE, and MAE, when  $Y_2$  is equal to 2003, 2006 and 2009, respectively. The ‘bold text’ indicates the best result. For two empirical datasets, our models achieve better prediction accuracy than the Q model. Specifically, we find that the IQ-2 model nearly always achieves the best accuracy, the IQ model yields slightly worse accuracy than the IQ-2 model, and the IQ-3 model gets lower accuracy than the IQ model. This indicates that the institutions in which scientists work, and the countries in which institutions are located are valuable prior information for scientific evaluation. Our model can effectively use the prior information to quantify the research ability of scientists and in turn, accurately predict their future impact. However, the inferior accuracy of the IQ-3 model compared with the IQ model may be due to the simplistic linear combination way of modelling author’s contribution to the paper, and a more sophisticated way for determining  $w_{\alpha, i}$  may be proposed in the future. Moreover, as shown in Figure 3, we report scatterplots of predicted and real h-index and  $C_{tot}$  of IQ-2 model when  $Y_2 = 2006$ , in which the x-axis represents the predicted value and the y-axis represents the real value. The error bar represents a 95% confidence interval. The dots were distributed around the diagonal line, which means that the IQ-2 model achieve good prediction accuracy. We also find similar results on the IQ and IQ-2 models, which are reported in our Supplemental Appendix 2.

To further show the better prediction performance of our models, we also report the change of prediction accuracy of the h-index and  $C_{tot}$  under  $Y_2 \in [2001, 2010]$ , as shown in Figures 4 and 5. The x-axis denotes  $Y_2$ , while the y-axis denotes the evaluation metrics. The brown, red, blue and black lines indicate the results of the Q, IQ, IQ-2 and IQ-3 models,

**Table 4.** Prediction performance of the h-index.

Y <sub>2</sub>	Metrics	Field	Q model	IQ model	IQ-2 model	IQ-3 model
2003	<i>Pearsonr</i>	CS	0.8337	0.8951	<b>0.8952</b>	0.8218
		Physics	0.8231	0.8784	0.8838	0.8435
	<i>R</i> <sup>2</sup>	CS	0.6151	0.6944	<b>0.7036</b>	0.5038
		Physics	0.5630	0.7574	0.7697	0.4104
	RMSE	CS	3.1605	2.8164	<b>2.7735</b>	3.5886
		Physics	3.4510	2.5711	2.5051	4.0086
MAE	CS	2.1862	1.9213	<b>1.8945</b>	2.7586	
	Physics	2.2275	1.6890	1.6871	3.0254	
2006	<i>Pearsonr</i>	CS	0.6694	0.7895	<b>0.7903</b>	0.6821
		Physics	0.5787	0.6960	0.7236	0.7127
	<i>R</i> <sup>2</sup>	CS	0.0218	0.4053	<b>0.4274</b>	0.2901
		Physics	0.0000	0.4293	0.4901	0.1182
	RMSE	CS	6.0636	4.7279	<b>4.6393</b>	5.1658
		Physics	6.5197	4.3717	4.1323	5.4343
MAE	CS	4.5026	3.1755	<b>3.1322</b>	4.1553	
	Physics	4.4842	2.7353	2.7219	4.4398	
2009	<i>Pearsonr</i>	CS	0.5833	0.7318	<b>0.7330</b>	0.5897
		Physics	0.4067	0.5643	0.5870	<b>0.6175</b>
	<i>R</i> <sup>2</sup>	CS	0.0000	0.2472	<b>0.2761</b>	0.2657
		Physics	0.0000	0.1418	0.1888	0.1379
	RMSE	CS	8.3517	6.0781	<b>5.9603</b>	6.0028
		Physics	9.2269	6.1511	5.9805	6.1653
MAE	CS	6.2949	4.0600	<b>4.0163</b>	4.6755	
	Physics	6.5506	3.8495	3.8378	4.8608	

IQ: institution Q; CS: computer science; RMSE: root mean square error; MAE: mean absolute error.

**Table 5.** Prediction performance of C<sub>tot</sub>

Y <sub>2</sub>	Metrics	Field	Q model	IQ model	IQ-2 model	IQ-3 model
2003	<i>Pearsonr</i>	CS	0.8099	0.8803	<b>0.8822</b>	0.8553
		Physics	0.4120	0.7203	0.8526	0.8035
	<i>R</i> <sup>2</sup>	CS	0.6359	0.7612	<b>0.7624</b>	0.6717
		Physics	0.0000	0.4730	0.7101	0.6423
	RMSE	CS	443.8554	359.4097	<b>358.5023</b>	421.4412
		Physics	951.7609	468.2211	347.2690	385.7627
MAE	CS	135.0992	116.7510	<b>115.9322</b>	147.9782	
	Physics	134.6501	113.9692	111.5963	159.6420	
2006	<i>Pearsonr</i>	CS	0.6746	0.8241	<b>0.8301</b>	0.7916
		Physics	0.2943	0.5124	0.7477	0.6935
	<i>R</i> <sup>2</sup>	CS	0.4000	0.6722	<b>0.6786</b>	0.5416
		Physics	0.0000	0.0752	0.5164	0.4753
	RMSE	CS	775.5136	573.1875	<b>567.6203</b>	677.8759
		Physics	1478.2727	897.7273	649.2014	676.2072
MAE	CS	282.2775	232.7257	<b>229.3469</b>	295.0730	
	Physics	299.5449	238.5675	233.6282	305.6894	
2009	<i>Pearsonr</i>	CS	0.5555	0.7823	<b>0.7887</b>	0.7350
		Physics	0.2298	0.4758	0.6544	0.4238
	<i>R</i> <sup>2</sup>	CS	0.1353	0.6060	<b>0.6139</b>	0.4527
		Physics	0.0000	0.1073	0.3706	0.3826
	RMSE	CS	1193.1132	805.4156	<b>797.3086</b>	949.2027
		Physics	2068.2193	1171.0453	983.2364	<b>973.8251</b>
MAE	CS	426.3945	341.2947	<b>337.1603</b>	432.3994	
	Physics	498.1137	389.5458	384.9858	464.9858	

IQ: institution Q; CS: computer science; RMSE: root mean square error; MAE: mean absolute error.

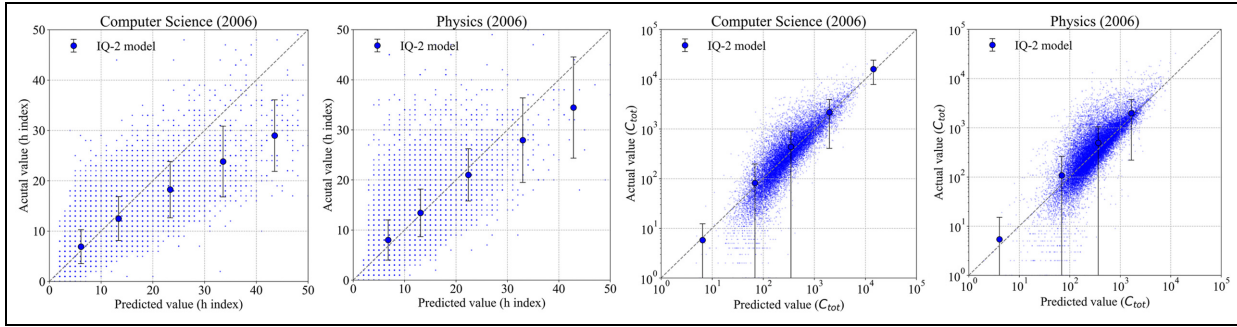


Figure 3. The prediction results of the IQ-2 when  $Y_2 = 2006$ .

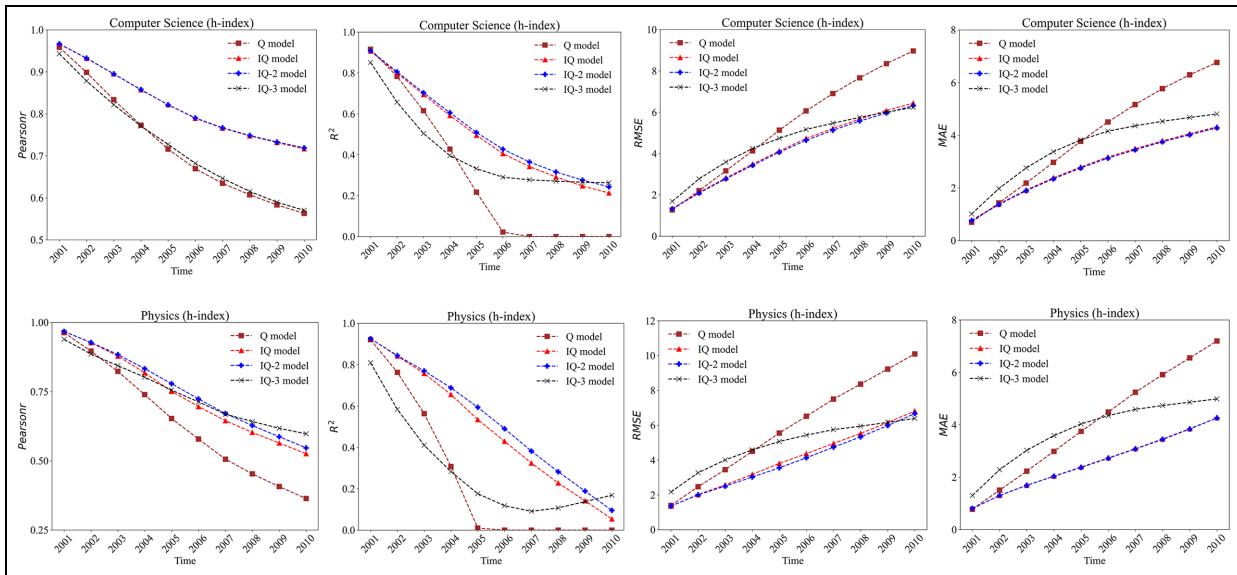


Figure 4. The prediction performance of the h-index when  $Y_2 \in [2001, 2010]$ .

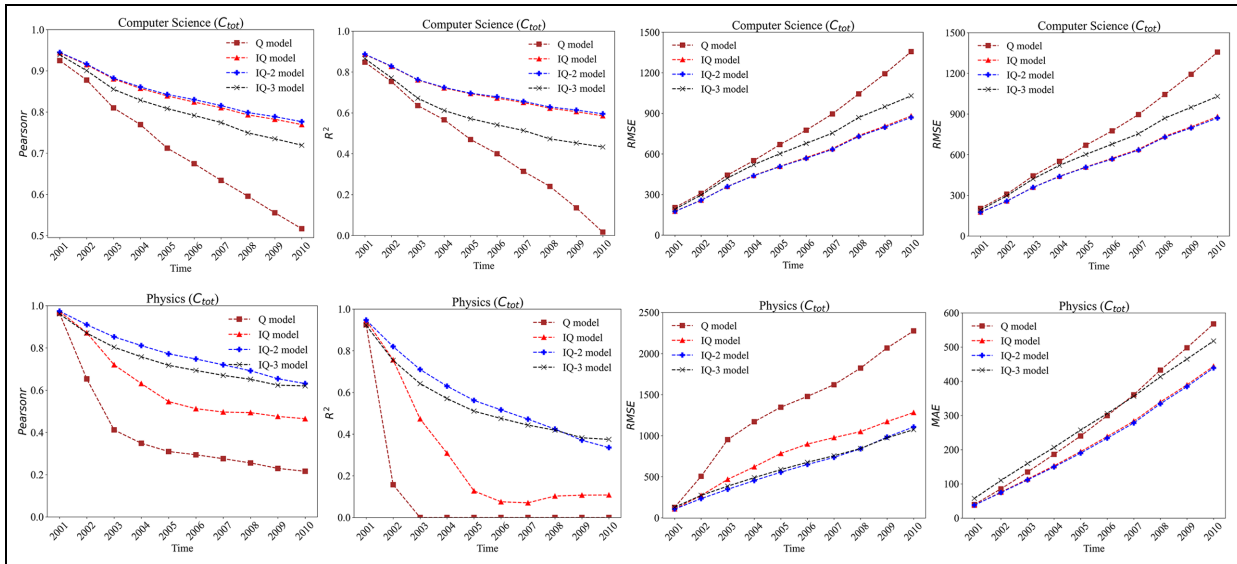
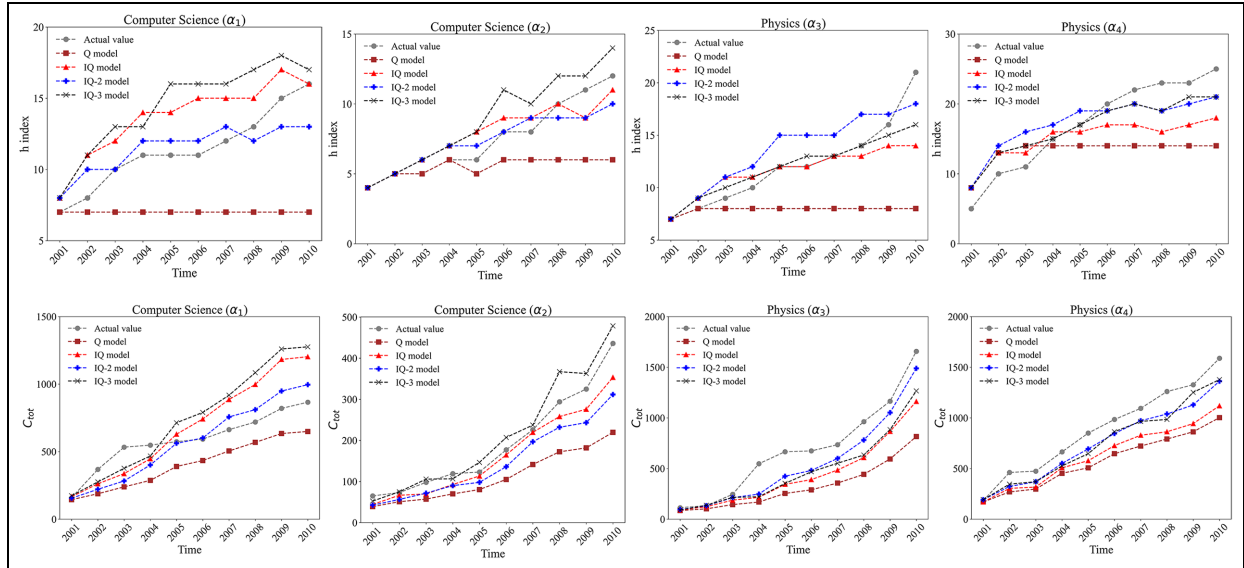


Figure 5. The prediction performance of  $C_{tot}$  when  $Y_2 \in [2001, 2010]$ .



**Figure 6.** Four case scientists in the prediction experiment.

respectively. We find that the IQ-2 model nearly always scores highest on *Pearsonr* and  $R^2$ , and scores lowest on *RMSE* and *MAE* under different  $Y_2$ . Overall, our models have more accurate prediction power than the Q model, which indirectly shows that they quantify research ability in a more accurate manner. In addition, with the increase of  $Y_2$ , the prediction accuracy of all models gradually decreases, and this trend is especially obvious in the Q model.

Finally, we also randomly select four scientists ( $\alpha_1, \alpha_2, \alpha_3$ , and  $\alpha_4$ ) as a case study to analyse our prediction results. As shown in Figure 6, we plot the h-index and  $C_{tot}$  as a function of  $Y_2$ . The grey line indicates the actual value, while brown, red, blue and black lines indicate the predicted value of the Q, IQ, IQ-2 and IQ-3 models, respectively. We find the blue line is closest to the grey line, followed by red and black lines. Thus, the IQ-2 model achieves the best prediction accuracy, and IQ, IQ-2 and IQ-3 models achieve better prediction performance than the Q model on the empirical data.

### 5.5. The prediction performance compared with common machine learning models

We compare our model with common machine learning models, which have been widely employed in predicting the scientific impact [18,51,52]. Specifically, we select Support Vector Regression (SVR) and Random Forest Regression (RFR) as representatives of traditional machine learning models, and Long Short Term Memory network (LSTM) and Gated Recurrent Network (GRU) as representatives of deep learning models. For each empirical dataset, scientists are split into the training set, verification set, and test set as a ratio of 8:1:1. We divide it 10 times and repeat the above experiments to predict the h-index and  $C_{tot}$ . The input of SVR, RFR, GRU and LSTM is a time series with a length of eleven (e.g.  $C_{tot}^1, C_{tot}^2, \dots, C_{tot}^{11}$ ), and the output is a scalar (e.g.  $C_{tot}^{12}$ ), in which the academic environment information is not included. The SVR and RFR are implemented based on the sklearn library, and the LSTM and GRU networks are implemented based on the Tensorflow framework. As shown in Tables 6 and 7, we report the average prediction performance on the test set. The bold text indicates the best result. We find that the IQ-2 model always achieves the best prediction performance in forecasting  $C_{tot}$ , and achieves competitive performance in forecasting the h-index than SVR, RFR, GRU and LSTM. Actually, one obvious advantage of machine learning models is that they can be seen as a ‘black box’ and are easy to deploy and universally applicable. However, our graphical models focus more on explainable power in scientific evaluation, and can quantify the hidden research ability.

## 6. Discussion

We present a series of practical guidelines for using the institution Q model, and summarise the theoretical and practical implications of this study. First,  $C_{\alpha,i}^{10}$  takes a long time to be observed. Previous studies have shown that the majority of papers acquire most citations within the first three years of publication [50], and citations acquired within three and five

**Table 6.** Prediction performance of the h-index.

Y <sub>2</sub>	Metrics	Field	RFR	SVR	GRU	LSTM	IQ-2
2003	Pearsonr	CS	0.8542	0.8650	0.8696	0.8695	<b>0.8936</b>
		Physics	0.8638	0.8708	0.8724	0.8723	0.8830
	R <sup>2</sup>	CS	0.7273	0.7355	0.7548	<b>0.7552</b>	0.7033
		Physics	0.7451	0.7471	0.7603	0.7600	<b>0.7692</b>
	RMSE	CS	2.6256	2.5859	2.4895	<b>2.2874</b>	2.7363
		Physics	2.6559	2.6454	2.5753	2.5746	<b>2.5275</b>
MAE	CS	1.9584	<b>1.8119</b>	1.8653	1.8464	1.8742	
	Physics	1.8662	1.7316	1.8019	1.8004	<b>1.6915</b>	
2006	Pearsonr	CS	0.7137	0.7398	0.7439	0.7427	<b>0.7903</b>
		Physics	0.6532	0.6775	0.6775	0.6785	0.7199
	R <sup>2</sup>	CS	0.5010	0.5254	<b>0.5507</b>	0.5488	0.4281
		Physics	0.4194	0.4341	0.4573	0.4580	<b>0.4843</b>
	RMSE	CS	4.3122	4.2059	<b>4.0923</b>	4.1006	4.6167
		Physics	4.4191	4.3633	4.2733	4.2698	<b>4.1640</b>
MAE	CS	3.2285	<b>2.9961</b>	3.0560	3.0387	3.1114	
	Physics	3.1328	2.9061	3.0058	3.0158	<b>2.7429</b>	
2009	Pearsonr	CS	0.6136	0.6493	0.6500	0.6496	<b>0.7317</b>
		Physics	0.4866	0.5288	0.5267	0.5250	0.5845
	R <sup>2</sup>	CS	0.3646	0.3907	<b>0.4207</b>	0.4204	0.2672
		Physics	0.2214	0.2424	0.2751	0.2749	0.1839
	RMSE	CS	5.5602	5.4464	<b>5.3097</b>	5.3112	5.9688
		Physics	5.8475	5.7683	5.6428	5.6471	5.9852
MAE	CS	4.1747	<b>3.8862</b>	3.9567	3.9797	4.0319	
	Physics	4.2117	3.8757	4.0242	4.0699	<b>3.8371</b>	

RFR: Random Forest Regression; SVR: Support Vector Regression; GRU: Gated Recurrent Network; LSTM: Long Short Term Memory network; IQ: institution Q; CS: computer science; RMSE: root mean square error; MAE: mean absolute error.

**Table 7.** Prediction performance of C<sub>tot</sub>.

Y <sub>2</sub>	Metrics	Field	RFR	SVR	GRU	LSTM	IQ-2
2003	Pearsonr	CS	0.8353	0.8305	0.8534	0.8372	<b>0.8771</b>
		Physics	0.8291	0.8044	0.7994	0.8021	0.8361
	R <sup>2</sup>	CS	0.6923	0.6552	0.6620	0.6675	<b>0.7533</b>
		Physics	0.6804	0.6270	0.6129	0.6198	0.6805
	RMSE	CS	403.2105	427.4737	422.9294	418.7360	<b>360.0250</b>
		Physics	375.8929	407.8437	415.4518	411.5571	377.1841
MAE	CS	152.1931	145.8666	148.0416	149.0748	<b>115.3052</b>	
	Physics	138.8131	142.2504	149.8415	150.1761	115.4793	
2006	Pearsonr	CS	0.7422	0.7378	0.7512	0.7506	<b>0.8358</b>
		Physics	0.7159	0.6942	0.6671	0.6770	0.7496
	R <sup>2</sup>	CS	0.5331	0.5019	0.5204	0.5287	<b>0.6862</b>
		Physics	0.4896	0.4292	0.3996	0.4115	0.5128
	RMSE	CS	683.0272	708.1028	694.2919	687.2806	<b>552.8815</b>
		Physics	665.4531	706.0731	724.5224	717.5693	653.2078
MAE	CS	301.3233	289.0176	290.0828	294.0581	<b>228.5504</b>	
	Physics	294.1592	293.7415	303.2461	301.5355	236.9563	
2009	Pearsonr	CS	0.6798	0.6909	0.6929	0.6868	<b>0.7763</b>
		Physics	0.5793	0.5528	0.5373	0.5451	0.6390
	R <sup>2</sup>	CS	0.4397	0.4181	0.4276	0.4122	<b>0.5930</b>
		Physics	0.3058	0.2503	0.2468	0.2467	0.3518
	RMSE	CS	954.7665	972.8313	964.3413	977.2964	<b>814.0659</b>
		Physics	986.9753	1026.9343	1029.5145	1029.1385	955.7702
MAE	CS	448.1728	434.5136	436.0893	443.3566	<b>338.5184</b>	
	Physics	470.7465	468.3011	480.6622	482.8770	381.5656	

RFR: Random Forest Regression; SVR: Support Vector Regression; GRU: Gated Recurrent Network; LSTM: Long Short Term Memory network; IQ: institution Q; CS: computer science; RMSE: root mean square error; MAE: mean absolute error.

years are an important reflection of a paper's quality [51–53]. Hence,  $C_{\alpha,i}^3$  and  $C_{\alpha,i}^5$  may be suitable choices to quantify the quality of a publication in a timely manner, and are used in the institution Q model to evaluate a scientist's research ability and predict their scientific impact in time. In addition, the citations count standardised by the z-score strategy for alleviating the time bias [54,55] is also a feasible alternative. In this article, to ensure the same experimental setup as utilised in the Q model, we still choose  $C_{\alpha,i}^{10}$ . Second, the BBVI-EM algorithm belongs to a stochastic optimisation algorithm, and therefore the initial value of the algorithm is very important. Practically, the  $Q_{\alpha}$  estimated by the Q model tends to be a good initial value, which is also adopted in this study. In addition, a series of training techniques applicable to the gradient descent algorithm (such as the learning rate decay, early stopping, data subsampling) can be utilised to ensure better performance of the BBVI-EM algorithm.

This article has the following theoretical implications. First, we present a novel institution Q model (IQ) and its two variants (IQ-2 and IQ-3), which integrate the academic environment (i.e. scientists' institutions, countries and collaborators) as valuable prior information, and jointly evaluate the research ability of scientists from different institutions. Second, our models achieve excellent quantification performance of research ability and satisfactory prediction performance of future scientific impact. Specifically, our models can more accurately estimate the hidden research ability of scientists and institutions than the Q model and can more accurately predict the h-index and  $C_{\text{tot}}$  of scientists than common machine learning models. The IQ-2 model gets the optimal results. The simple idea of taking academic environment as prior knowledge does not surprisingly improve the performance of our models. The potential reason is that our models can flexibly use the information of other scientists from an institution to evaluate the research ability of a scientist from the institution. Hence, our models can better cope with the data scarcity problem frequently encountered in scientific research evaluation. Third, the idea behind our models broadens the thought for incorporating academic environment into scientific evaluation. Other factors related to the academic environment, possibly affecting scientists' research performance, may also be utilised as priori information, and other researchers can draw lessons from our modelling methods. Hence, this study makes a theoretical contribution to guide how to integrate academic environment into the evaluation of scientific activities. Finally, we propose a generic BBVI-EM algorithm, which achieves excellent performance on our issue.

The article also has the following practical implications. First, the excellent estimation performance and the satisfactory prediction performance of our models suggest that they are not only effective evaluation tools to quantify the research ability of scientists but also practical prediction tools to predict scientists' scientific impact. Thus, funding agencies may adopt our model to improve their talent assessment mechanism, evaluate and rank the research ability of scientists, and target promising young scholars to stimulate scientific innovation [56]. Moreover, the BBVI-EM algorithm is a general variational inference algorithm, which can also be used in the inference and estimation of other probabilistic graphical models. The BBVI-EM algorithm inherits the merits of the BBVI algorithm and allows researchers to easily explore a wide variety of models without suffering from the tedious derivation process of probabilistic graphical models.

## 7. Conclusion

In this study, we propose the institution Q model with an explainable generative process to comprehensively consider the academic environment, research ability and randomness in the citation process, which can effectively quantify the research ability of scientists and institutions. To approximate the posterior distribution in our models, we also present a universal and effective BBVI-EM algorithm. We examine the quantification performance of research ability and prediction performance of the scientific impact of our models on synthetic data and empirical data, which shows that our models are effective evaluation and prediction tools for scientific evaluation.

There are still some limitations in this study. Our models focus on using scientists' institutions, countries and collaborators as prior information. However, there are many other factors which may be used as valuable information for research evaluation, such as research topics, gender, and co-authorship network [14,28,57]. Hence, a more sophisticated model including as much comprehensive prior information as possible needs to be explored. Moreover, previous studies have shown that author name order, author contribution statement and contribution list represent the contribution of authors to a publication [4]. Therefore, this information may be used to meticulously design  $w_{\alpha,i}$  in the IQ-3 model in the future.


## Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.


## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by the Youth Science Foundation of the National Natural Science Foundation of China (grant no. 72004168).

## ORCID iDs

Shengzhi Huang  <https://orcid.org/0000-0002-7035-4627>

Wei Lu  <https://orcid.org/0000-0002-0929-7416>

Zhuoran Luo  <https://orcid.org/0000-0003-0677-8350>

## Supplemental material

Supplemental material for this article is available online.

## Note

1. <https://github.com/WannaLearning/Quantifying-scientists-research-ability-by-taking-institutions-scientific-impact-as-priori-informa.git>.

## References

- [1] Zhou Y, Wang R, Zeng A et al. Identifying prize-winning scientists by a competition-aware ranking. *J Informetr* 2020; 14: 101038.
- [2] Meng Q and Kennedy PJ. Discovering influential authors in heterogeneous academic networks by a co-ranking method. In: *Proceedings of the 22nd ACM international conference on information & knowledge management*, 2013, pp. 1029–1036, <https://dl.acm.org/doi/abs/10.1145/2505515.2505534>
- [3] Simoes N and Crespo N. A flexible approach for measuring author-level publishing performance. *Scientometrics* 2020; 122: 331–355.
- [4] Yang S, Xiao A, Nie Y et al. Measuring coauthors' credit in medicine field – based on author contribution statement and citation context analysis. *Inf Process Manag* 2022; 59: 102924.
- [5] Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci* 2005; 102: 16569–16572.
- [6] Jiang J, Shi P, An B et al. Measuring the social influences of scientist groups based on multiple types of collaboration relations. *Inf Process Manag* 2017; 53: 1–20.
- [7] Liu X, Bollen J, Nelson ML et al. Co-authorship networks in the digital library research community. *Inf Process Manag* 2005; 41: 1462–1480.
- [8] Martín-Martín A, Orduña-Malea E and López-Cózar ED. Author-level metrics in the new academic profile platforms: the online behaviour of the Bibliometrics community. *J Informetr* 2018; 12: 494–509.
- [9] Mikki S, Zygmuntowska M, Gjesdal ØL et al. Digital presence of Norwegian scholars on academic network sites – where and who are they? *PLoS One* 2015; 10: e0142709.
- [10] Sinatra R, Wang D, Deville P et al. Quantifying the evolution of individual scientific impact. *Science* 2016; 354: aaf5239.
- [11] Li W, Zhang S, Zheng Z et al. Untangling the network effects of productivity and prominence among scientists. *Nat Commun* 2022; 13: 4907.
- [12] Way SF, Morgan AC, Larremore DB et al. Productivity, prominence, and the effects of academic environment. *Proc Natl Acad Sci* 2019; 116: 10729–10733.
- [13] Abbasi A, Altmann J and Hossain L. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *J Informetr* 2011; 5: 594–607.
- [14] Bu Y, Ding Y, Xu J et al. Understanding success through the diversity of collaborators and the milestone of career. *J Assoc Inform Sci Technol* 2018; 69: 87–97.
- [15] Lungeanu A, Huang Y and Contractor NS. Understanding the assembly of interdisciplinary teams and its impact on performance. *J Informetr* 2014; 8: 59–70.
- [16] Zhang F, Liu X, Tang J et al. OAG: toward linking large-scale heterogeneous entity graphs. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2585–2595, <https://dl.acm.org/doi/10.1145/3292500.3330785>
- [17] Xie Q, Zhang X and Song M. A network embedding-based scholar assessment indicator considering four facets: research topic, author credit allocation, field-normalized journal impact, and published time. *J Informetr* 2021; 15: 101201.
- [18] Huang S, Huang Y, Bu Y et al. Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Inf Process Manag* 2022; 59: 102799.
- [19] Bormmann L and Daniel H-D. Convergent validation of peer review decisions using the h index: extent of and reasons for type I and type II errors. *J Informetr* 2007; 1: 204–213.



- [20] Bornmann L, Wallon G and Ledin A. Is the h index related to (standard) bibliometric measures and to the assessments by peers? An investigation of the h index by using molecular life sciences data. *Res Eval* 2008; 17: 149–156.
- [21] Lovegrove BG and Johnson SD. Assessment of research performance in biology: how well do peer review and bibliometry correlate? *Bioscience* 2008; 58: 160–164.
- [22] Rousseau R and Leuven K. Reflections on recent developments of the h-index and h-type indices. *Collnet J Scientometr Inf Manag* 2008; 2: 1–8.
- [23] Egghe L. Theory and practise of the g-index. *Scientometrics* 2006; 69: 131–152.
- [24] Alonso S, Cabrerizo F, Herrera-Viedma E et al. Hg-index: a new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics* 2010; 82: 391–400.
- [25] Radicchi F, Fortunato S, Markines B et al. Diffusion of scientific credits and the ranking of scientists. *Phys Rev E* 2009; 80: 056103.
- [26] Senanayake U, Piraveenan M and Zomaya AY. The p-index: ranking scientists using network dynamics. *Procedia Comput Sci* 2014; 29: 465–477.
- [27] Bioglio L, Rho V and Pensa RG. Ranking by inspiration: a network science approach. *Mach Learn* 2020; 109: 1205–1229.
- [28] Jiang J, Shi P, An B et al. Measuring the social influences of scientist groups based on multiple types of collaboration relations. *Inf Process Manag* 2017; 53: 1–20.
- [29] Haustein S, Peters I, Bar-Ilan J et al. Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics* 2014; 101: 1145–1163.
- [30] Ortega JL. Relationship between altmetric and bibliometric indicators across academic social sites: the case of CSIC’s members. *J Informetr* 2015; 9: 39–49.
- [31] Fang Z, Costas R, Tian W et al. How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *J Assoc Inform Sci Technol* 2021; 72: 918–932.
- [32] Lemke S, Mazarakis A and Peters I. Conjoint analysis of researchers’ hidden preferences for bibliometrics, altmetrics, and usage metrics. *J Assoc Inform Sci Technol* 2021; 72: 777–792.
- [33] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
- [34] Colombo-Mendoza LO, Valencia-García R, Rodríguez-González A et al. Towards a knowledge-based probabilistic and context-aware social recommender system. *J Inform Sci* 2018; 44: 464–490.
- [35] Gershman SJ, Blei DM, Norman KA et al. Decomposing spatiotemporal brain patterns into topographic latent sources. *Neuroimage* 2014; 98: 91–102.
- [36] Kendall A and Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*. Curran Associates Inc., Red Hook, NY, USA, 2017; 30: 5580–5590.
- [37] Manning JR, Zhu X, Willke TL et al. A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. *Neuroimage* 2018; 180: 243–252.
- [38] Simonetti A, Albano A, Plaia A et al. Ranking coherence in topic models using statistically validated networks. *J Inform Sci*. Epub ahead of print 20 January 2023. DOI: 10.1177/01655515221148369.
- [39] Ranganath R, Gerrish S and Blei D. Black box variational inference. In: *Proceedings of the 7th international conference on artificial intelligence and statistics*, 2013, pp. 814–822, <https://oar.princeton.edu/handle/88435/pr15v50>
- [40] Rathore AS and Roy D. Performance of LDA and DCT models. *J Inform Sci* 2014; 40: 281–292.
- [41] Bishop CM and Nasrabadi NM. *Pattern recognition and machine learning*. New York: Springer, 2006.
- [42] Griffiths TL and Steyvers M. Finding scientific topics. *Proc Natl Acad Sci* 2004; 101: 5228–5235.
- [43] Rosen-Zvi M, Griffiths T, Steyvers M et al. The author-topic model for authors and documents. arXiv preprint arXiv:1207.4169, 2012, <https://arxiv.org/abs/1207.4169>
- [44] Mandt S, Hoffman MD and Blei DM. Stochastic gradient descent as approximate Bayesian inference. arXiv preprint arXiv:1704.04289, 2017, <https://arxiv.org/abs/1704.04289>
- [45] Zhao X, Wang D, Zhao Z et al. A neural topic model with word vectors and entity vectors for short texts. *Inf Process Manag* 2021; 58: 102455.
- [46] Ning X, Zheng Y, Jiang Z et al. Nonparametric topic modeling with neural inference. *Neurocomputing* 2020; 399: 296–306.
- [47] Pluchino A, Biondo AE and Rapisarda A. Talent versus luck: the role of randomness in success and failure. *Adv Complex Syst* 2018; 21: 1850014.
- [48] Sobkowicz P, Frank RH, Biondo AE et al. Inequalities, chance and success in sport competitions: simulations vs empirical data. *Phys A Stat Mech Appl* 2020; 557: 124899.
- [49] Van Hooydonk G. Fractional counting of multiauthored publications: consequences for the impact of authors. *J Am Soc Inform Sci* 1997; 48: 944–945.
- [50] Wang D, Song C and Barabási A-L. Quantifying long-term scientific impact. *Science* 2013; 342: 127–132.
- [51] Abrishami A and Aliakbary S. Predicting citation counts based on deep neural network learning techniques. *J Informetr* 2019; 13: 485–499.
- [52] Ruan X, Zhu Y, Li J et al. Predicting the citation counts of individual papers via a BP neural network. *J Informetr* 2020; 14: 101039.

- 
- [53] Cui Y, Wang Y, Liu X et al. Multidimensional scholarly citations: characterizing and understanding scholars' citation behaviors. *J Assoc Inform Sci Technol* 2023; 74: 115–127.
- [54] Gao Q, Liang Z, Wang P et al. Potential index: revealing the future impact of research topics based on current knowledge networks. *J Informetr* 2021; 15: 101165.
- [55] Guan J, Yan Y and Zhang JJ. The impact of collaboration and knowledge networks on citations. *J Informetr* 2017; 11: 407–422.
- [56] Shang J, Zeng M and Zhang G. Investigating the mentorship effect on the academic success of young scientists: an empirical study of the 985 project universities of China. *J Informetr* 2022; 16: 101285.
- [57] Duch J, Zeng XHT, Sales-Pardo M et al. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One* 2012; 7: e51332.