https://doi.org/10.51408/issi2025_070

Analysis of Relationships Between Paper Citations and Their Category Influencing Factors: A Bayesian Network with Latent Variables Approach

Mingyue Sun[1], Mingliang Yue [2], Wen Peng[3], Tingcan Ma[4]

[1]*sunmingyue22@mails.ucas.ac.cn*, [2]*yueml@mail.whlib.ac.cn*, [3]*pengwen23@mails.ucas.ac.cn*, [4]*matc@whlib.ac.cn*

Chinese Academy of Sciences, National Science Library (Wuhan), 430071 Wuhan (China)
University of Chinese Academy of Sciences, Department of Information Resources Management, School of Economics and Management, 100190 Beijing (China)

Abstract

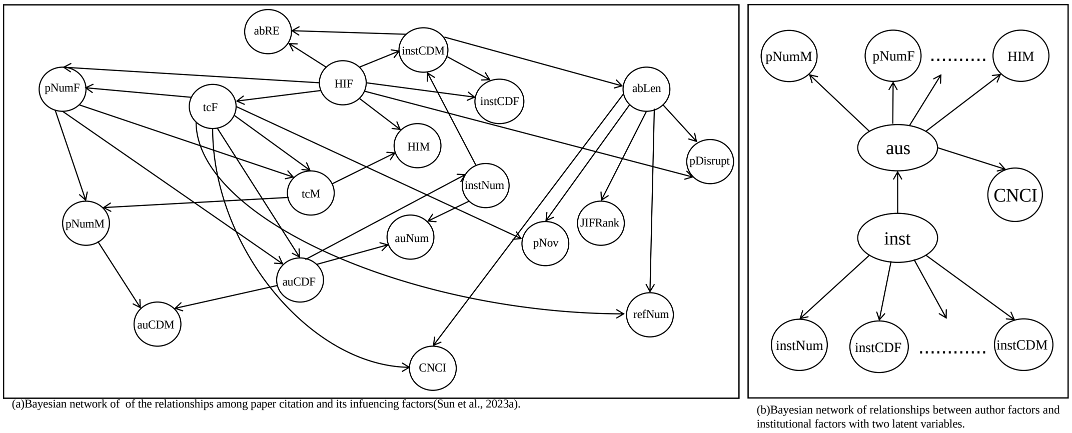
The analysis of the impact of academic papers has long been a topic of interest among scholars. Many studies have been carried out to explore the interaction between paper citations and its influencing factors from a microscopic perspective, e.g., analyzing the correlation between individual or multiple observable variables (such as author h index and publication counts) and citation count. However, it remains challenging to conduct analysis from a relatively macroscopic perspective, such as understanding how author characteristics as a whole influence citation count. In this paper, we adopt a Bayesian Network (BN) with latent variables as the knowledge framework, using latent variables to describe characteristics of different aspects (i.e., institution, author and paper aspects) as a whole, so that interactions among latent category factors as well as observable factors can be analyzed. We use the K-means algorithm to acquire categories of latent variables and use constraint-based scoring approach to learn the BN. We analyzed how the introduction of latent variables provides new perspectives compared to using only observable variables, conducted corresponding analyses, and reached certain conclusions.

Introduction

Citation plays a crucial role in the scientific evaluation of publications, individual scientists, and research institutions, prompting the academic community to contemplate the mechanisms and rationale behind its use for evaluation purposes. Numerous scholars have studied the factors that influence citation rates and how they affect the number of citations (Bornmann, 2011; Xie et al., 2019).

According to Tahamtan and Bornmann (2018a), the process of a research paper being cited is complex. There are significant relationships between paper citations and various factors (Xie et al., 2019), including authorship characteristics such as academic influence, gender, academic background, and others (Hurley et al., 2013; Ruan et al., 2020; Stremersch et al., 2015; Wang et al., 2019a, 2019b), as well as institutional and/or national affiliations (Didegah & Thelwall, 2013). Other influential factors include the impact of the publishing journal (Bornmann & Leydesdorff, 2015; Stegehuis et al., 2015), linguistic properties of the paper such as readability (Didegah & Thelwall, 2013; Stremersch et al., 2015), the paper's innovativeness (Wu et al., 2019), the number and impact of references (Bornmann & Leydesdorff, 2015), and other considerations like scientific funding (Rigby, 2013) and open access status (McCabe & Snyder, 2014), among others.

While previous studies have analyzed the independent or joint associations between various factors and paper citations, there has been relatively insufficient consideration of the correlations between these influencing factors. Sun et al. (2023) addressed this gap by applying Bayesian network (BN) to study the interactive relationships between citation and its influencing factors, utilizing 20 variables. However, the constructed network structure may be complex. As depicted in Figure 1(a), the intricate dependency relationships represented by directed edges between nodes may hinder effective focus on specific analyses of interest, such as understanding how author factors as a whole influence the citation impact of papers. To simplify the BN structure and facilitate more intuitive inference, latent variables can be introduced (Koller & Friedman 2009, Zhang & Guo 2006). For instance, the introduction of latent variables, represented as *aus* (author factors) and *inst* (institutional factors), significantly streamlined the dependency relationships between variables, as shown in Figure 1(b). Meanwhile, as demonstrated in the Results section, this streamlined network can allows for new analytical perspectives.



(a)Bayesian network of of the relationships among paper citation and its infuencing factors(Sun et al., 2023a).

(b)Bayesian network of relationships between author factors and institutional factors with two latent variables.

Figure 1. Bayesian Network (with latent variables).

Therefore, this paper adopts a Bayesian network incorporating latent variables (as categorical factors) influencing paper citations, including author factor, institutional factor, and factor related to paper-specific characteristics. To differentiate from traditional BN models that directly use observable variables (Sun et al., 2023) , this study applies a domain-specific latent variable learning method. This method captures implicit patterns across multiple observable variables, enabling the BN structure to reflect higher-level macroscopic interrelations between latent variables with reduced complexity.The BN structure is learned based on a constraint-based scoring algorithm that incorporates domain expert knowledge. After modeling, BN inferences are conducted to discover new analytical perspectives and draw conclusions.

The remainder of this paper is structured as follows: Section 2 introduces the necessary knowledge of Bayesian networks with latent variables. Section 3 outlines the construction process of the BN, including optimal structure learning and parameter learning of BN with latent variables. Section 4 demonstrates BN inference and presents findings. Section 5 concludes the paper.

Bayesian network with latent variables

A Bayesian network with latent variables (Zhang & Guo, 2006) is defined as a binary tuple (G, θ). $\mathsf{G} = (\chi, \mathsf{E})$ represents a DAG structure, where the node set $\chi = \{\mathsf{x}_1, \ldots, \mathsf{x}_\mathsf{n}\}$ consists of each node corresponding to a random variable, $\square = \square \cup \square$ indicates that $\square$ includes both the observable variable set $\square$ and the latent variable set $\square$, with $|\square| + |\square| = \square$、 $|\square| > 0$ □□□ $|\square| \geq 0$. $\square$ represents the set of directed edges, where a directed edge $\square_\square \rightarrow \square_\square$ indicates a dependency relationship from node $\square_\square$ to node $\square_\square$, or a causal relationship where $\square_\square$ is a direct cause of $\square_\square$. θ represents the set of conditional probability parameters, denoted as $\pi(\square_\square) = \{\square_\square \mid < \square_\square, \square_\square > \in \mathsf{E}\}$. If all nodes are discrete variables, $\square_\square = \{\square(\square_\square|\square(\square_\square))\}$ represents the conditional probability distribution (CPT) of node $\square_\square$, and $\square_{\square\square\square} = \{\square(\square_\square = \square|\square(\square_\square) = \square)\}$ represents the conditional probability parameter corresponding to the situation where node $\square_\square$ takes on value k and its parent nodes take on the jth combination of values. In Figure 1(b), the set of latent variables are {*aus*, *inst*}, and the set of observable variables is {*pNumM*, *pNumF*, *HIM*, *instCDM*, *instCDF*, *instNum*, ...}.

The construction of Bayesian network with latent variables mainly consists of four parts: determining the number of latent variables, determining the cardinality of latent variables, structure learning, and parameter learning (Koller & Friedman, 2009; Zhang & Guo, 2006), as demonstrated in Fig. 2. Determining the number of latent variables involves deciding how many latent variables are needed in the model. Methods for this include clustering-based approaches (Wang et al., 2008; Mourad et al., 2013) and clique-based methods (Elidan et al., 2000; He et al., 2014). Determining the cardinality of latent variables refers to determining the number of states each latent variable can take. Typically, clustering techniques are employed, treating latent variables as hidden categories. The process involves starting with a small number of categories (e.g., binary) and incrementally increasing them until the objective function reaches a maximum, with the corresponding category number considered as the cardinality of latent variables (Elidan & Friedman, 2013). In practice, the number and cardinality of latent variables are highly domain-specific and often determined by experts after analyzing the scenario (Wu & Yue, 2023). Structure learning aims to find the optimal network structure that fits the real data best using scoring-based search (Chickering, 2002; Ramsey et al., 2017; Goudet et al., 2018; Zhu et al., 2019) or conditional independence evaluation (Kong & Wang, 2023; Colombo & Maathuis, 2014). Parameter learning algorithms commonly utilize the EM algorithm and its variants (Qi et al., 2022; Kan et al., 2022).

In the network, latent variables often serve as abstractions of multiple observable variables, capturing the combined effects of the observable variables. Therefore, latent class model is usually adopted as local structure to model the relationships among latent variables and their corresponding observable variables. In the latent class model, observable variables are only connected to their corresponding latent class variables, and do not connect with other variables (Zhang & Guo, 2006). The latent variables and other (latent or observable) variables can be interconnected to form a global network, thereby establishing relationships between the represented observable variables and other variables. The structure learning process is then used to determine the global structure, based on certain domain-specific constraints (Yue et al., 2020).

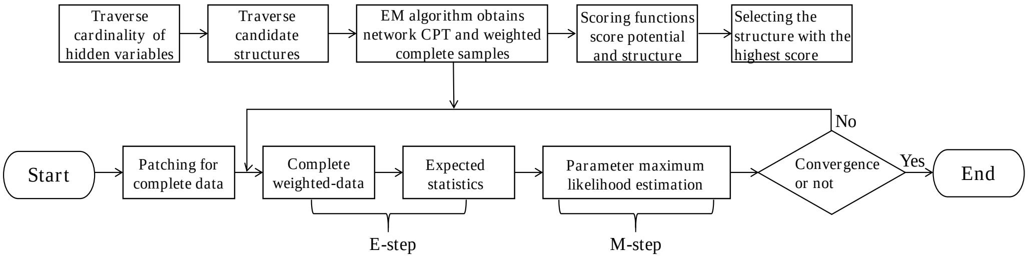

Traverse cardinality of hidden variables
Traverse candidate structures
EM algorithm obtains network CPT and weighted complete samples
Scoring functions score potential and structure
Selecting the structure with the highest score
No
Start
Patching for complete data
Complete weighted-data
Expected statistics
Parameter maximum likelihood estimation
Convergence or not
Yes
End
E-step
M-step

Figure 2. The flowchart for learning the structure and parameters of a Bayesian network with latent variables.

Bayesian Network construction

In this section, we first introduce the latent and observable variables considered in the BN. Then we discuss the constraints on the global structure based on the nature of academic citation. Finally, we present the BN construction algorithm.

The latent variables and corresponding observable variables

Unlike Sun et al., (2023) that focus on analyzing individual observable variables, this paper introduces three latent variables—paper factors, inst_factors and aus_factors—to abstract and integrate diverse observable variables into higher-level macroscopic dimensions. The rationale for selecting these latent variables is grounded in the citation mechanism outlined by Tahamtan and Bornmann (2018), which highlights the multifaceted influences on a paper's ability to garner citations. According to their findings, the intrinsic value of a research paper is a key determinant of its academic influence, while author characteristics significantly shape the citing author's expectations of the document's value. These author characteristics are further categorized into Author-level factors and Platform-level factors, both of which are posited to influence the perceived value of a paper. Building on this understanding, this study identifies three latent variables—paper factors, inst_factors and aus_factors.This latent variables simplifies the representation of complex observable variables interactions while preserving critical dependencies of paper itself, authors and institutes.

Paper_factors represents a latent variable that integrates multiple observable variables (characteristics) of the paper's research content, which collectively capture the paper's academic value. *Paper_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, such as novelty (pnov) (Bu et al. 2021) and the number of references and the citations they received (refNum, refcitation_sum, refcitation_average) (Rigby 2013; Onodera and

Yoshikane 2015; Xie et al. 2019; Bornmann and Leydesdorff 2015). Each category represents a distinct level of academic value, reflecting the combined influence of these related observable variables. These related obesrvable variables encompasses the novelty (pnov) (Bu et al. 2021) and disruptiveness (pDisrupt) (Wu et al. 2019) of a paper, reflecting aspects of a research work's contribution. The number of references and the citations they received (refNum, refcitation_sum, refcitation_average) (Rigby 2013; Onodera and Yoshikane 2015; Xie et al. 2019; Bornmann and Leydesdorff 2015) reflect the amount of knowledge and impact of knowledge referenced by the work, as well as linguistic properties influencing other researchers' understanding of the paper. This includes text readability (abER) (Stremersch et al. 2015; Lei and Yan 2016; Ante 2022) and text length (abstract_length, title_length, key_words_length) (Vamclay 2013; Xie et al. 2019; Ruan et al. 2020; Stremersch et al. 2015).

aus_factors represents a latent variable that integrates multiple observable variables (characteristics) related to the impact of authors of the research papers. *aus_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, such as the number of papers published (pNum_Max, pNum_average, pNumF) (Stremersch et al. 2015), the number of citations received by published papers (tc_Max, tc_average, tcF) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022). Each category represents a distinct level of the combined impact of the first authors and corresponding authors in a research paper, reflecting the combined influence of these related observable variables.These related obesrvable variables encompasses the number of papers published (pNum_Max, pNum_average, pNumF) (Stremersch et al. 2015), the number of citations received by published papers (tc_Max, tc_average, tcF) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022), the h-index (h_max, h_average, HIF) (Wang et al. 2012; Wang et al. 2019; Xie et al. 2019), centrality measures in the collaboration network (auCDF, degree_max, degree_average), eigenvector centrality (Eigenvector_centrality_Max, Eigenvector_centrality_F, Eigenvector_centrality_average) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021), and the number of authors per paper (authors) (Yu et al. 2014; Bornmann and Leydesdorff 2015; Xie et al. 2019).

inst_factors represents a latent variable that integrates multiple observable variables related to the institutions influence of the paper authors. *inst_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, including centrality measures of research institutes in the collaboration network (inst_degree_average, inst_degree_max, inst_degree_F) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021), eigenvector centrality of research institutes (inst_Eigenvector_centrality_average, inst_Eigenvector_centrality_max, inst_Eigenvector_centrality_F) (Zhang et al. 2021), and the number of research institutes (institution) (Wang et al. 2019; Zhang et al. 2021). Each category represents a distinct level of the combined impact of the institutions affiliated with the authors in a research paper, reflecting the combined influence of these related observable variables. Table 1 presents the latent variables and their corresponding observable variables.

Finally, research work affects the paper quality, we utilize Normalized Citation Impact (CNCI) to evaluate the quality of academic papers. This is because, as Li (2019) pointed out, current academic paper evaluation methods mainly characterize from the perspectives of impact and innovativeness. According to Tahamtan et al. (2016), creativity and novelty are features influencing internal factors of papers, and we classify paper innovativeness into *paper_factors*. Based on the extensive use of CNCI by scholars in the field of scientometrics to measure paper impact, and the fact that conclusions based on this metric are generally considered representative (Lei and Yan 2016; Ante 2022; Bornmann and Leydesdorff 2015), this paper only employs CNCI to evaluate the quality of academic papers.

Table 1. Latents Variable and their Corresponding Observable Variables in a BN.

Latent variables	*corresponding observable variables*	*meaning of corresponding observable variables*
paper_ factors	pnov	Paper novelty (Bu et al. 2021)
	pDisrupt	Paper disruption (Wu et al. 2019)
	refsNum	Number of references (Rigby 2013; Onodera and Yoshikane 2015; Xie et al. 2019)
	abER	Summary text readability (Stremersch et al. 2015; Lei and Yan 2016; Ante 2022)
	abatract_length	Summary text length (Vamclay 2013; Xie et al. 2019; Ruan et al. 2020)
	title_length	Title text length (Stremersch et al. 2015; Xie et al. 2019)
	key_words_length	keyword text length (Xie et al. 2019)
	refcitation_average	Average number of citations for references (Bornmann and Leydesdorff 2015; Xie et al. 2019)
	refcitation_sum	Total number of citations of references (Xie et al. 2019)
	Rank	CCF Rank (Qian et al. 2017)
aus_factors	pNum_Max	Number of published papers (max) (Stremersch et al. 2015)
	pNum_average	Number of published papers (average) (Stremersch et al. 2015)
	pNumF	Number of published papers (first author) (Stremersch et al. 2015)
	tcF	Total citations (first author) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022)
	tc_Max	Total citations (max) (Xie et al. 2019; Amjad et al. 2022)
	tc_average	Average citations (Xie et al. 2019; Amjad et al. 2022)
	HIF	h-index (first author) (Wang et al. 2012; Wang et al. 2019; Xie et al. 2019）
	h_max	h-index (max) (Hurley et al. 2013; Xie et al. 2019)
	h_average	h-index (average) (Xie et al. 2019)
	auCDF	Co-authorship network centrality degree (first author) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	degree_max	Co-authorship network centrality degree (max) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	degree_average	Co-authorship network centrality degree (average) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)

	Eigenvector_centrality_Max	Co-authorship network eigenvector centrality (max) Co-authorship network eigenvector centrality (first author) Co-authorship network eigenvector centrality (average) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	Eigenvector_centrality_F	
	Eigenvector_centrality_average	
	authors	Number of authors (Yu et al. 2014; Bornmann and Leydesdorff 2015; Xie et al. 2019)
inst_factors	institution	Number of institutes (Wang et al. 2019; Zhang et al. 2021)
	inst_degree_average	Cooperation network centrality degree (institute with average value) Cooperation network centrality degree (institute with maximum value) Cooperation network centrality degree (institute with first author value) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	inst_degree_max	
	inst_degree_F	
	inst_Eigenvector_centrality_average	Cooperation network eigenvector centrality (institute with average value) Cooperation network eigenvector centrality (institute with maximum value) Cooperation network eigenvector centrality (institute with first author value) (Zhang et al. 2021)
	inst_Eigenvector_centrality_max	
	inst_Eigenvector_centrality_F	

Structural Constraints

The potential structures of the BN are illustrated in Figure 3, consisting of two parts: local structure and global structure. The local structure is the latent class model mentioned earlier. The structure constraints for global network include: (1) Authors and institutions may be able to reference each other; (2) Authors and institutions can reference paper features, but not vice versa; (3) Authors, institutions, and paper features may be able to reference CNCI, but not vice versa. Given these constraints, there are a total of 58 compliant potential network structures. Our goal is to use a BN learning algorithm to select the structure that best matches the data distribution and estimate the parameters accordingly. Fig. 3 demonstrates a potential network structure. The *aus_factors* directly influences (points to) the *paper_factors*, *inst_factors*, and *CNCI*. Similarly, the *inst_factor*, in turn, directly influences (points to) the *aus_factors*, *paper_factors*, and *CNCI*. Moreover, the *paper_factors* directly influences (points to) *CNCI*.

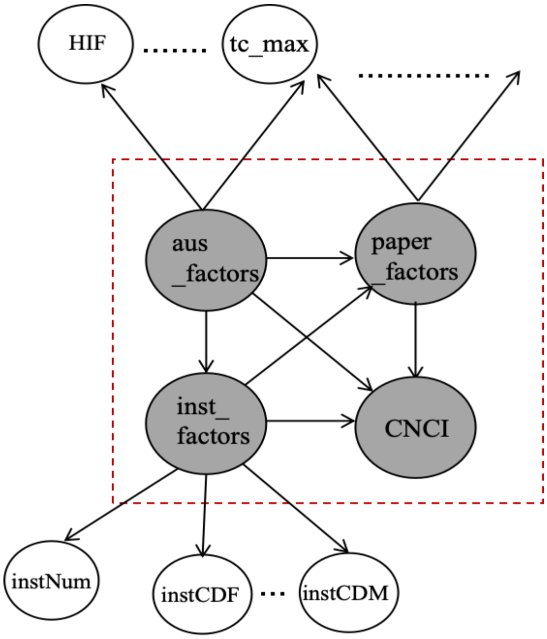


Figure 3. A potential network structure.

Bayesian Network (BN) with latent variables construction

Data preparation

The paper utilizes the Aminer paper dataset (Tang, 2008) as the foundational data, which is employed for computing the observable variables in Table 1. The Aminer dataset comprises a comprehensive collection of academic research papers and citation relationships, and it has been extensively utilized in various research endeavors related to academic research evaluation (Abramo et al., 2019; Amjad et al., 2022; Shao et al., 2022; Song et al., 2018). It has been employed in numerous studies associated with academic research evaluation. The dataset provides detailed information including paper identification number (*id*), title (*title*), publication date (*year*), author details (including identification numbers (*_id*), names (*name*), institutional affiliations (*org*), and institutional identification numbers (*gid*)), publication venues (including publication identification numbers (*_id*), publication names (*raw*)), keywords, abstract, citation counts (*n_itation*), reference number, and complete citation relationships among papers. Based on this information, we can compute the values of all the listed variables in Table 1 except for CCF Rank (https://www.ccf.org.cn/c/2019-12-01/666146.shtml). Regarding CCF Rank, given that our dataset covers academic journals and conferences in the field of computer science, and considering the significant influence of CCF rankings along with the absence of metrics such as JIF for conference papers, we introduce CCF Rank as a substitute for JIF Rank. This approach aids in accurately reflecting the importance of the papers.

In addition to CCF Rank, prior to BN learning, the factor values should be discretized into states. The values of CCF Rank can be A, B, or C. The discretization rule for other factors utilizes the equal-width binning method, whereby variables are sorted in ascending order according to numerical values and divided into four equal intervals.

As shown in Table 2, we give the reasons for the missing values of various factors in the Aminer data. For conducting latent variable class learning based on K-Means, all attributes (i.e., the variables in Table 1) must have values, which requires each record to be complete. Hence, 96,760 complete records are used as the data source for the BN learning.

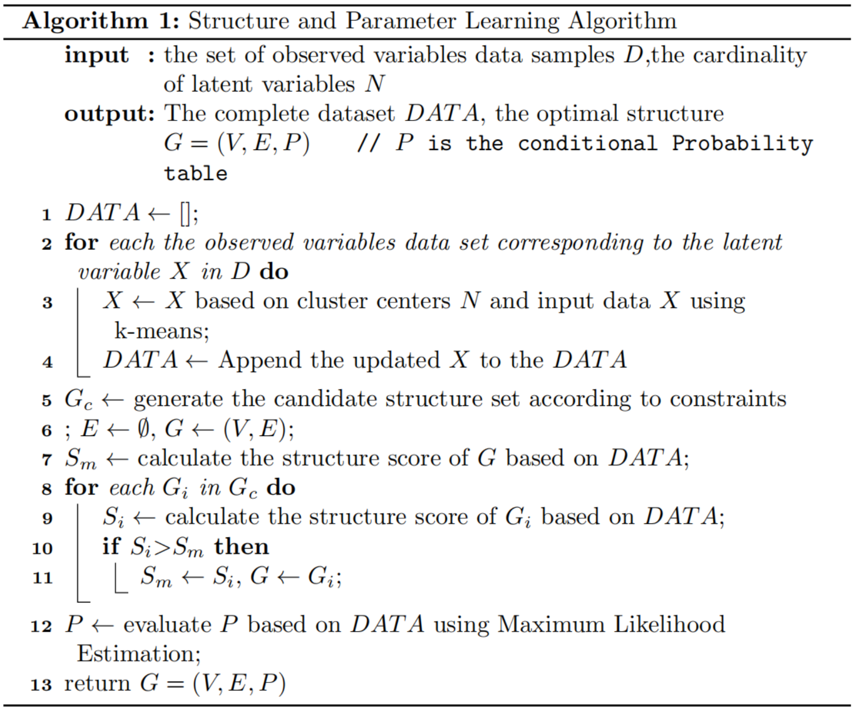
Table 2. Reasons for missing Factors values.

Factors	Missing reasons
auCDF, Eigenvector_centrality_F, HIF, tcF, pNumF	The first author identification number (First_aus_id) is missing
authors,degree_max,degree_average, Eigenvector_centrality_Max Eigenvector_centrality_averge h_max, h_average, tc_Max, tc_average pNum_Max, pNum_average	The entire author field is missing
inst_degree_F, inst_Eigenvector_centrality_F	The first author's institution identification number is missing
institution, inst_degree_average, inst_degree_max inst_Eigenvector_centrality_average inst_Eigenvector_centrality_max	Author's institution field is missing
abER, abatract_length	Summary field missing
title_length, key_words_length	Reference field missing
refsNum, refcitation_average, refcitation_sum, pnov, pDisrupt	(1) Reference field is missing. (2) Lack of real reference relationships
CNCI	The number of citations field is missing
Rank	Lack of publications. Only the grades (A, B, C) of journals and conferences in the CCF catalog are retained in publications.

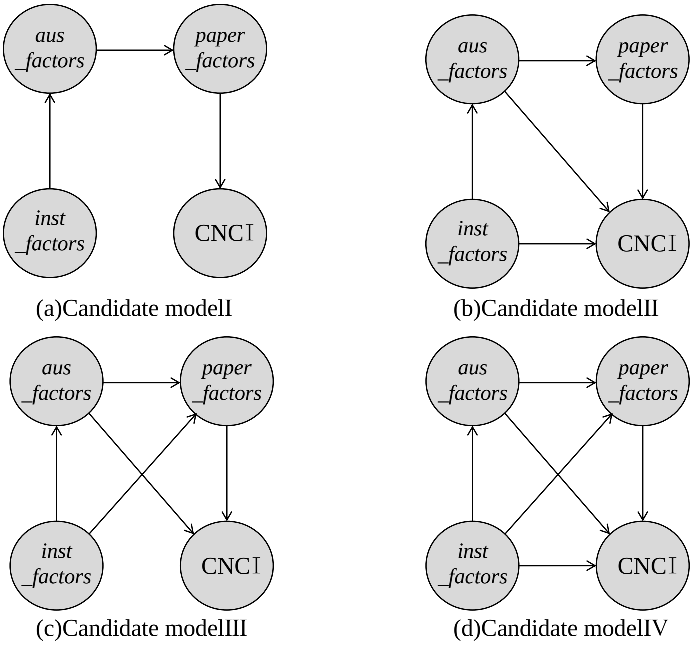
Learning algorithm

Based on the given data, we employ the BN learning algorithm to learn its optimal structure and parameters. The input data consists of the observable variables $D = [aus, inst, paper]$, where *aus* represents the list of observable variables corresponding to the latent variable *aus_factors*, *inst* represents the list of observable variables corresponding to the latent variable *inst_factors*, and *paper* represents the list of observable variables corresponding to the latent variable *paper_factors*. The cardinality of the latent variables is determined to be 4 based on expert knowledge (drawing on journal classification). The output includes the complete dataset as well as the optimal structure and its parameters. First, the algorithm uses the K-means algorithm to cluster the observable variables, obtaining categories corresponding to the latent variables (line 1-4). Next, all possible candidate structures are generated based on the structure constraints (line 5). Then, a scoring function is used to evaluate these candidate structures to find the optimal structure (line 6-11). Finally, the corresponding parameters for the optimal candidate structure are calculated (line 12-13). We implemented Algorithm 1 using the sklearn package(https://scikit-learn.org) and the pgmpy package(https://pgmpy.org/). The sklearn package covers almost all mainstream machine learning algorithms. It provides wrappers for common machine learning algorithms, including classification, regression, clustering, and dimensionality reduction. pgmpy is a pure python implementation for the BN with a focus on structure learning, parameter estimation, approximate and exact inference.The data preparation procedures (data preprocessing, variable value calculation and discretization) were also implemented in Python. Since the K-means clustering algorithm requires a complete dataset to learn the latent variable categories, some data values may be unavailable due to missing data. Therefore, we removed the data with missing values and used the complete dataset to learn the latent variable categories. The Bayesian Information Criterion (BIC) (Schwarz, 1978) is used as the scoring metric to evaluate whether a candidate model is suitable for a given dataset. According to the structural constraints given above, we obtained a total of 58 candidate structure sets. As shown in Figure 4, we present some candidate structure sets. The structure with the highest score is considered the optimal structure. Once the optimal structure is determined, the network parameters can be easily learned from the data using the Maximum Likelihood Estimation (Zhang & Guo, 2006).

In the end, we obtained two optimal structures as shown in Figure 5. This model suggests that the reputation of institutions (*inst_factors*), the capability/influence of authors (*aus_factors*), and the features of the papers (*paper_factors*) all have impact on *CNCI*, and there is no single factor that can isolate the influence of another factor on the *CNCI*. Further, the results indicate that there is no explicit directional relationship between the influence of authors and institutions, meaning it is not clear whether the influence of authors determines the influence of institutions, or vice versa. Furthermore, the two optimal models are Markov equivalent (Zhang & Guo, 2006), which means they share the same probabilistic implications. Therefore, in subsequent analyses, as shown in Figure 5, optimal model (a) will be employed for inference and analysis. The learned BN with latent variables is shown in Figure 6.



```
Algorithm 1: Structure and Parameter Learning Algorithm
    input : the set of observed variables data samples $D$,the cardinality
            of latent variables $N$
    output: The complete dataset $DATA$, the optimal structure
            $G = (V, E, P)$    // $P$ is the conditional Probability
            table
$DATA \leftarrow []$;
for each the observed variables data set corresponding to the latent
  variable $X$ in $D$ do
    $X \leftarrow X$ based on cluster centers $N$ and input data $X$ using
      k-means;
    $DATA \leftarrow$ Append the updated $X$ to the $DATA$
$G_c \leftarrow$ generate the candidate structure set according to constraints
; $E \leftarrow \emptyset$, $G \leftarrow (V, E)$;
$S_m \leftarrow$ calculate the structure score of $G$ based on $DATA$;
for each $G_i$ in $G_c$ do
    $S_i \leftarrow$ calculate the structure score of $G_i$ based on $DATA$;
    if $S_i > S_m$ then
        $S_m \leftarrow S_i$, $G \leftarrow G_i$;
$P \leftarrow$ evaluate $P$ based on $DATA$ using Maximum Likelihood
  Estimation;
return $G = (V, E, P)$
```



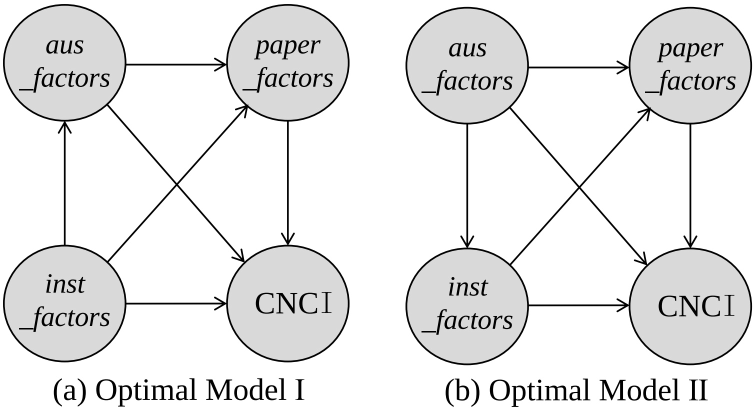
(a)Candidate modelI

(b)Candidate modelII

(c)Candidate modelIII

(d)Candidate modelIV

Figure 4. Some candidate models.



(a) Optimal Model I

(b) Optimal Model II

Figure 5. Optimal models.

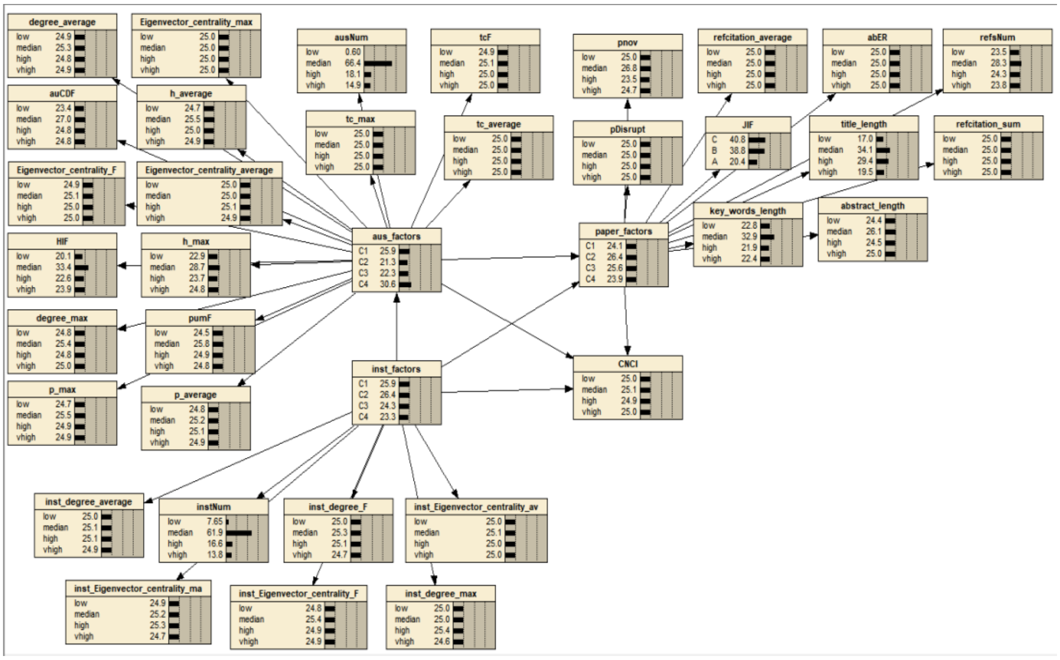
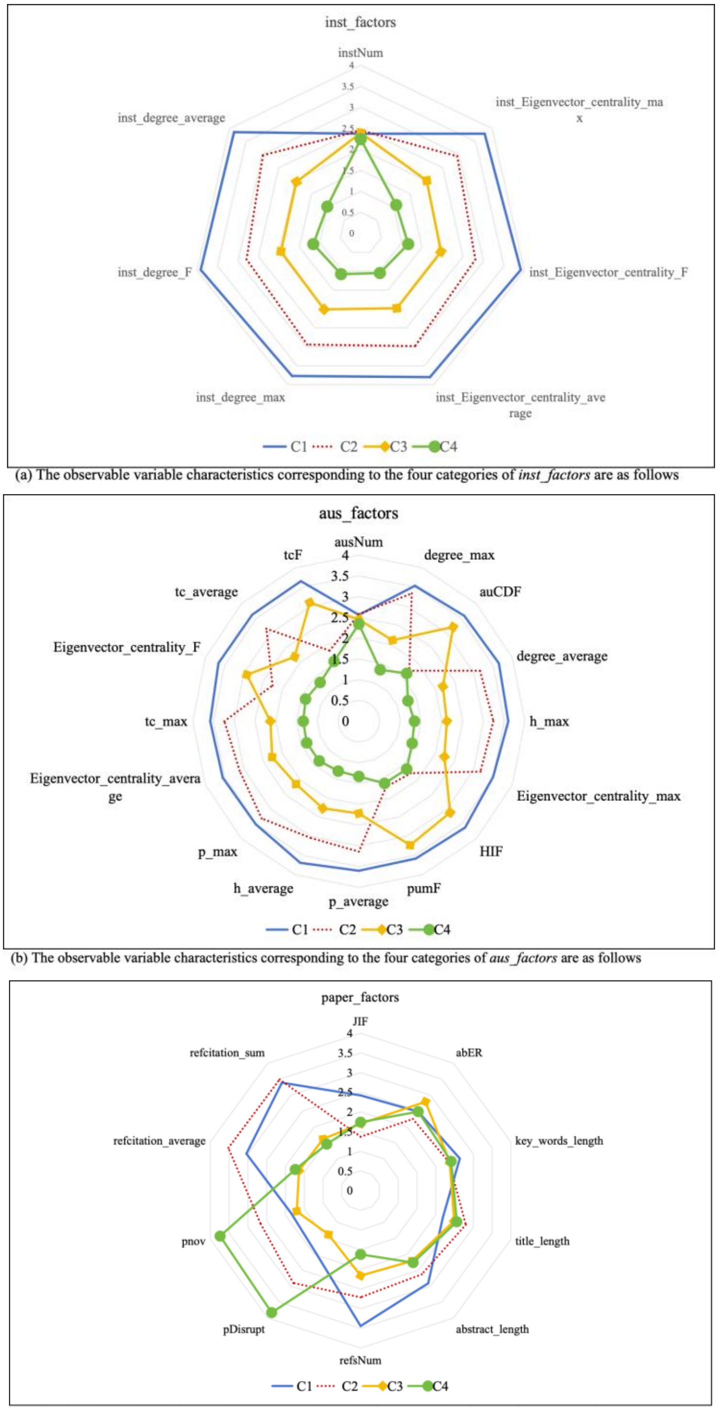

degree_average low 24.9 median 25.3 high 24.8 vhigh 24.9
Eigenvector_centrality_max low 25.0 median 25.0 high 25.0 vhigh 25.0
ausNum low 0.60 median 66.4 high 18.1 vhigh 14.9
tcF low 24.9 median 25.1 high 25.0 vhigh 25.0
pnov low 25.0 median 26.8 high 23.5 vhigh 24.7
refcitation_average low 25.0 median 25.0 high 25.0 vhigh 25.0
abER low 25.0 median 25.0 high 25.0 vhigh 25.0
refsNum low 23.5 median 28.3 high 24.3 vhigh 23.8
auCDF low 23.4 median 27.0 high 24.8 vhigh 24.8
h_average low 24.7 median 25.5 high 25.0 vhigh 24.9
tc_max low 25.0 median 25.0 high 25.0 vhigh 25.0
tc_average low 25.0 median 25.0 high 25.0 vhigh 25.0
pDisrupt low 25.0 median 25.0 high 25.0 vhigh 25.0
JIF C 40.8 B 38.8 A 20.4
title_length low 17.0 median 34.1 high 29.4 vhigh 19.5
refcitation_sum low 25.0 median 25.0 high 25.0 vhigh 25.0
Eigenvector_centrality_F low 24.9 median 25.1 high 25.0 vhigh 25.0
Eigenvector_centrality_average low 25.0 median 25.0 high 25.1 vhigh 24.9
key_words_length low 22.8 median 32.9 high 21.9 vhigh 22.4
abstract_length low 24.4 median 26.1 high 24.5 vhigh 25.0
aus_factors C1 25.9 C2 21.3 C3 22.3 C4 30.6
paper_factors C1 24.1 C2 26.4 C3 25.6 C4 23.9
HIF low 20.1 median 33.4 high 22.6 vhigh 23.9
h_max low 22.9 median 28.7 high 23.7 vhigh 24.8
degree_max low 24.8 median 25.4 high 24.8 vhigh 25.0
pumF low 24.5 median 25.8 high 24.9 vhigh 24.8
inst_factors C1 25.9 C2 26.4 C3 24.3 C4 23.3
CNCI low 25.0 median 25.1 high 24.9 vhigh 25.0
p_max low 24.7 median 25.5 high 24.9 vhigh 24.9
p_average low 24.8 median 25.2 high 25.1 vhigh 24.9
inst_degree_average low 25.0 median 25.1 high 25.1 vhigh 24.9
instNum low 7.65 median 61.9 high 16.6 vhigh 13.8
inst_degree_F low 25.0 median 25.3 high 25.1 vhigh 24.7
inst_Eigenvector_centrality_av low 25.0 median 25.1 high 25.0 vhigh 25.0
inst_Eigenvector_centrality_ma low 24.9 median 25.2 high 25.3 vhigh 24.7
inst_Eigenvector_centrality_F low 24.8 median 25.4 high 24.9 vhigh 24.9
inst_degree_max low 25.0 median 25.0 high 25.4 vhigh 24.6

Figure 6. The learned BN with Latent Variables of Model (a).

Results

Based on the four categories (C1, C2, C3, C4) of the three latent variables (*inst_factors*, *aus_factors*, *paper_factors*), we calculated the mean values of the corresponding observable variables, as shown in Figure 7. Intuitively, one might expect a certain partial order relationship among the average values of the categories of the observable variables, allowing us, for example, to determine when comparing two categories of authors that one category is superior to another (in a certain sense). However, as illustrated in Figure 7(c), there is an overlap between categories C2 and C3 in terms of the average values of the observable variables corresponding to the four categories of *aus_factors*. This overlap arises from the fact that the data used to learn the BN is paper oriented. Papers are often authored by a group of scholars with varying characteristics (such as h-index), and authors from different clusters may exhibit certain intersections in terms of variable values. For instance, in a highly influential paper authored by three scholars, the h-index of each author might appear as (high, high, high), (high, high, low), or (high, medium, low), among others. That is, high-impact papers are not necessarily co-authored solely by high-impact authors, and similarly, low-impact papers may not be co-authored solely by low-impact authors. This scenario leads to the overlap between categories C2 and C3. The same situation also occurs in the *paper_factors* in Fig. 5(b). The latent variables in this paper are used to describe the overall influences of the categories of the corresponding observable variables as a whole, where these categories are learned from the combinations of the real situations implied in the real bibliometrics data.



(a) The observable variable characteristics corresponding to the four categories of *inst_factors* are as follows

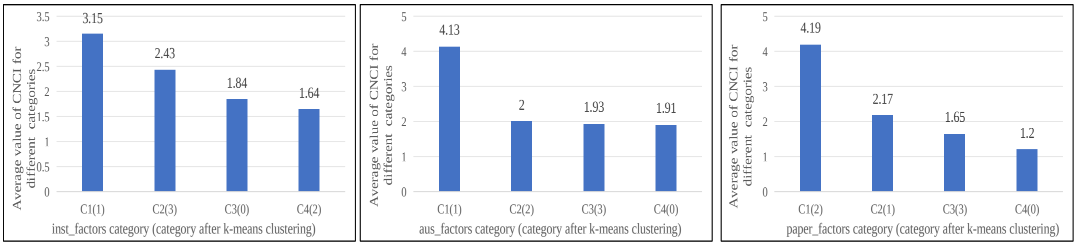
(b) The observable variable characteristics corresponding to the four categories of *aus_factors* are as follows

(c) The observable variable characteristics corresponding to the four categories of *paper_factors* are as follows

Figure 7. Differential characteristics presented by the four categories of latent variables in terms of observable variables.

To gain an understanding of the influence of *inst_factors*/*aus_factors*/*paper_factors* in each category, we calculated the average CNCI for papers corresponding to each category, as shown in Figure 8. Taking *inst_factors* as an example, as shown in

Figure 8(a), categories 1, 3, 0, 2 correspond to average *CNCI* values of 3.15, 2.43, 1.84, 1.64, respectively. To indicate varying influence among the categories and for ease of subsequent analysis, we named each category based on the ranking of their average CNCI values, with lower-numbered categories representing higher influence. Therefore, for *inst_factors*, categories 1, 3, 0, and 2 are named C1, C2, C3, and C4 respectively. Figure 8(b) and (c) show the situations of *aus_factors* and *paper_factors*.



(a) Average value of CNCI different inst_factors categories (b) Average value of CNCI different aus_factors categories (c) Average value of CNCI different paper_factors categories

Figure 8. Average CNCI values corresponding to the four categories of latent variables.

The necessity of adding latent variables

Using the approach taken by Sun et al. (2023), which sets different state combinations of each observed variable, for example, in the case of observable variables related to the authors, we can represent the situation of the authors in the paper and observe how these combinations affect the CNCI of the paper. However, although the combination of authors in some papers is different, the academic impact of the papers they publish is very similar. Setting observed variables can represent that a certain type of author combination can publish papers with high or low academic impact, but it is difficult to simultaneously represent the effect of several types of author combinations in producing similar academic impact (such as higher or lower academic impact). For example, when we set pNumF=*low* and h_max=*low*, the probability of CNCI from *low* to *vhigh* is 36.8%, 27.7%, 21.1%, 14.1%; when we set pNumF=*median* and h_max=*low*, the probability of CNCI from *low* to *vhigh* are 36.3%, 27.6%, 21.3%, and 14.7%. This shows that the academic influence of papers published by these two types of author combinations is similar, but it is impossible to express these two types of author combinations at the same time by setting observed variables. Furthermore, when the number of observed variables involved in author combinations is large, The situation will become more complicated, (such as setting auCDF, Eigenvector_centrality_F, HIF, tcF, pNumF, authors,degree_max,degree_average, Eigenvector_centrality_Max,Eigenvector_centrality_averge, h_max, h_average, tc_Max, tc_average, pNum_Max, pNum_average at the same time). Furthermore, setting different state combinations of each observed variable (Sun et al., 2023) fails to capture the combinations of author characteristics (impact) in actual, paper-oriented scenarios. As noted in the first paragraph of the Results section, we manually set author's h-index to the (*high, high, high*) state, representing the expected combination of author characteristics that would produce papers with higher average impact (as indicated by the papers' average CNCI value). This assumption stems from the intuitive belief that a combination of (*high, high, high*) h-index values among authors is more likely to result in a paper with higher average impact. However, this method still fails to identify the combinations of author characteristics that contribute to producing papers with a higher average impact(as indicated by the papers' average CNCI value).

Using our approach,which sets a single latent variable, we can simplify the complex combination of observed variables and classify different combinations of authors that produce similar academic impact into the same category. As shown in Figure 8(b), taking *aus_factors* as an example, we divide the latent variables into 4 levels (C1, C2, C3, C4). Compared with the *aus_factors* of the C2, C3, and C4 categories, the *aus_factors* of the C1 category include various author combinations. What these author combinations have in common is that their published papers have the highest average academic impact. Moreover, the author characteristic combinations represented by each category of *aus_factors* is paper-oriented and reflects real scenarios. This fundamentally differs from the method in Sun et al. (2023), which represents the expected combinations of author characteristics. As stated in the first paragraph of the Result section, setting *aus_factors*=C1 captures combinations such as (*high, high, high*), (*high, high, low*), or (*high, medium, low*) in terms of the authors' h-indices. This reflects the actual author combinations in real papers and is paper-oriented. The latent variable helps clarify the author characteristic combinations that contribute to papers with higher average impact (as indicated by the papers' average CNCI value). This approach is fundamentally different from the method in Sun et al. (2023), which involves manually setting each author's h-index to (*high, high, high*) to represent the expected combination of author characteristics for producing papers with higher average impact (as indicated by the papers' average CNCI value).

Therefore, there is a distinction between the meanings of observable and latent variables. For example, in the case of authors, observable variables refer to authors with different levels of influence, such as those measured by h-index or the number of published papers, whereas latent variable(*aus_factors*) represents combinations of author characteristics corresponding to different papers average impact levels(as measured by the papers' average CNCI values). As shown in Figure 7(c), within the four categories of *aus_factors*, there is overlap between C2 and C3 in terms of the average values of observable variables. This non-hierarchical overlap, observed from the data perspective, suggests that author characteristic combinations corresponding to different paper average impact levels (as measured by the papers' average CNCI values) exhibit differences when compared to authors with varying levels of influence (ranging from *vhigh* to *low*).

Due to the differences in the meanings of observable and latent variables, it is clear that studying the interactions between observable variables differs significantly from studying the interactions between latent variables. For instance, in research involving institutions and authors, by jointly setting different states of observable variables (e.g., HIF = *high*, pNumF = *high*) and observing the distribution of institutions(e.g.,instCDM), interactions between observable variables typically focus on how high-impact authors are distributed across institutions with different reputations. In contrast, by individually setting different states of latent variables (e.g., *aus_factors*=C1) and observing the distribution of *inst_factors,* interactions between latent variables tend to adopt a paper-oriented perspective, focusing on how author characteristic combinations in papers with higher average impact (as measured by the papers' average CNCI value) are distributed across institutional characteristic combinations with varying levels of paper average impact (as measured by the papers' average CNCI value).

Furthermore, latent and observable variables differ in how they contribute to understanding the interaction between paper impact and the factors that influence paper impact. For instance, by jointly setting different states of observable variables (e.g., HIF=*high*, pNumF=*high*) and observing the distribution of paper impact, this approach focuses on the paper impact distribution of papers written by high-impact authors. By individually setting different states of latent variables (e.g., *aus_factors*=C1) and observing the distribution of paper impact, this method focuses on the influence distribution of papers written by author characteristics combinations with higher average paper impact (measured by the average CNCI value of the papers). This distinction reflects the differing research perspectives and methodologies of observable and latent variables in paper impact analysis. In general, the interactions between latent variables and their relationship to paper impact differ significantly from the role of interactions between observable variables in influencing paper impact. The following section, "***Inferring the BN with Latent Variables***," will provide an example from the data perspective, analyzing the differences between observable and latent variables and exploring how interactions between latent variables affect paper impact.

Finally, by introducing latent variables, this method, compared to Sun et al. (2023), enables the study of interactions between latent and observable variables. By analyzing these interactions, it is possible to reveal the characteristic combinations of authors at different levels of paper average impact across the observable variable dimensions. The following section, ***Characteristics of Different Categories of the Latent Variables***, provides a more detailed analysis.

In summary, the introduction of latent variables not only simplifies complex combinations of observed variables, helping to classify author/institution/paper itself combinations with similar academic impact into the same category, but also represents a real, paper-oriented combination of multiple author/institution/paper characteristics. Compared to the method of Sun et al. (2023), this approach better captures the interactions between latent variables in real, paper-oriented scenarios, as well as the interaction between these latent variables and paper impact. Furthermore, the introduction of latent variables allows for the study of the interactions between latent and observed variables, revealing the characteristics of latent variables at different levels of paper average impact across various observed variable dimensions. This expands our understanding of Analysis of relationships between paper citations and their category influencing factors, which enhances the depth of the research in higher-level macroscopic perspectives.

Characteristics of different categories of the latent variables

Now let's take a detailed look at characteristics of different categories of the latent variables, and explore the relationship between these characteristics and CNCI. First, from the perspective of *inst_factors* clustering, as shown in Figure 7(a), from C4 *inst_factors* to C1 *inst_factors*, the degree of collaboration within the institution (*inst_ degree_F*, *inst_ degree_Max*) and the importance of the institution in the network (*inst_Eigenvector_centrality_F*, *inst_Eigenvector_centrality_max*) increase. At the same time, the average degree of collaboration within the organization (*inst_degree_average*) and the average importance of the organization in the network (*inst_Eigenvector_centrality_average*) also increase. However, there is almost no significant difference in the number of institutions in the 4 categories of *inst_factors*. This suggests that academic work is more likely to be cited when all participating authors are from institutions with higher degrees of collaboration and greater importance within the collaborative network.

Then, from the perspective of clustering based on *aus_factors*, as shown in Figure 7(b), the mean value of each observable variable in C1 *aus_factors* is the highest. In contrast, C4 *aus_factors* has the lowest mean value for each observable variable. In C2 and C3 *aus_factors*, the corresponding author has a higher mean value for each observable variable in C2, while in C3, the first author has a higher mean value for each observable variable. This indicates that academic work is more likely to be cited when all co-authors in a paper exhibit high level of each observable variable.

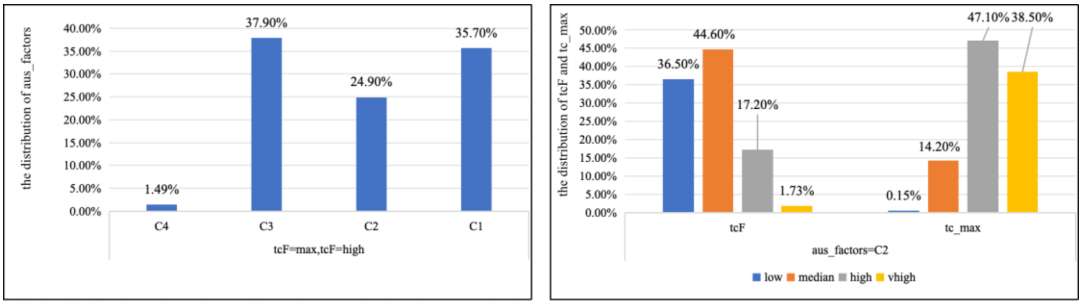
Further, from the perspective of clustering based on *paper_factors*, as shown in Figure 7(b), C1 *paper_factors* is generally published in the most influential publications. The number of references is the largest, and the average number of citations per reference and total number of citations of references are at a medium level. This shows that the research foundation of C1 *paper_factors* is relatively deep. In addition, C1 *paper_factors* tends to have the longest abstracts, the keywords with the largest number of words, and the most concise titles. However, C1 *paper_factors* is at a medium level of innovation and disruption. The number of references of C2 *paper_factors* is at a medium level, and the average number of citations per reference and total number of citations of references are the highest, indicating a deeper research foundation. Additionally, C2 *paper_factors* tends to have medium-length abstracts and the longest titles with the smallest number of keywords. It also exhibits moderate levels of innovation and disruption. Despite this, C2 *paper_factors* was published in the lowest impact journals. The number of references of C3 *paper_factors* is at a medium level, the average number of citations of references is at a medium level, the total number of citations is the lowest, and it lacks a deep research foundation. Additionally, C3 *paper_factors* tends to have the shortest abstracts, medium-length titles, and medium-level keywords. C3 *paper_factors* has the lowest level of innovation and disruption. C4 *paper_factors* is characterized by the highest level of disruption and innovation. This shows that high-impact works tend to be published in the highest-impact publications, with mid-range number of references, number of citations per reference, and total number of citations of references. They typically have the longest abstracts, the highest number of keywords, the most concise titles, and demonstrate a moderate level of innovation.

Finally, taking an overall perspective, hidden variables (*aus_factors*, *inst_factors*, *paper_factors*) possess their own characteristics, with some features having high impact while others have relatively lower impact. Through the analysis above, we conclude that in academic papers, (1) the higher the degree of collaboration of its institutional portfolio, the more important it is in the collaboration network and (2) the higher the influence of its author portfolio, the easier it is to be cited, in addition, (3) Additionally, the paper itself should be published in highly influential publications. It should have a moderate number of references, citations per reference, and total citations for references. Furthermore, it should feature a lengthy abstract, an extensive list of keywords, a succinct title, and display a moderate degree of innovation.

Inferring the BN with latent variables

This section will provide an example from the data perspective, analyzing the differences between observable and latent variables, observing the distributions of these three latent variables, and exploring how interactions between latent variables affect paper impact.

First, we will provide an example from the data perspective, analyzing the differences between observable and latent variables. Taking authors as a case study, we use observable variables tcF, tc_max to represent the number of citations of papers published by the first author and the corresponding author. We use the latent variable *aus_factors* to represent the combinations of author characteristics that result in different paper average impact levels (as measured by the papers' average CNCI values), as shown in Figure 9(a). When tcF=*high* and tc_max=*high*, the probability distribution of aus_factors from C4 to C1 is 1.49%, 37.90%, 24.90% and 35.70%. This indicates that authors with the same level of influence are not all categorized into the same group that produces papers with similar levels of average impact (as measured by the papers' average CNCI values). As shown in Figure 9(b), when aus_factors=C2, the probability distributions of tcF and tc_max from *low* to *vhigh* are 36.5%,44.60%, 17.20%,1.73% and 0.15%, 14.20%, 47.10%, 38.50% respectively. It can be observed that the probability distributions of tc_max and tcF from *low* to *vhigh* are not confined to a single state (i.e., the probability distribution is not 100% in one state). This indicates that the C2 category of *aus_factors* cannot be represented by a single joint setting of different states for tc_max and tcF. The setting of aus_factors=C2 is because the four levels of categories in aus_factors, namely C1, C2, C3, and C4, correspond to the four states of the observed variables: *vhigh*, *high*, *median*, and *low*. Through this example, it is clear that observable variables cannot represent latent variables through joint settings, and the meanings represented by latent variables are significantly different from those of observable variables.

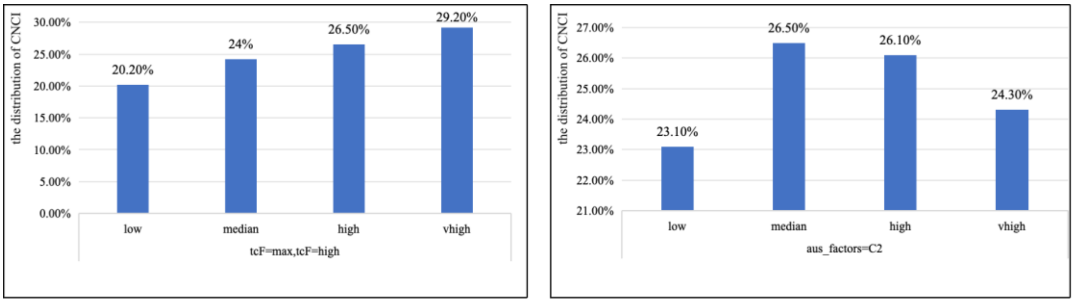


(a) the distribution of *aus_factors* by setting tcF=*high* and tc_max=*high*

(b) the distribution of tcF, tc_max by setting *aus_factors*=C2

Figure 9. The distribution of aus_factors, tcF and tc_max.

It is evident that the interactions between observable variables and CNCI differ meaningfully from those between latent variables and CNCI. As shown in Figure 10(a), when setting tcF = high and tc_max = high, the probability distribution of CNCI is 20.20%, 24%, 26.50%, and 29.20%, which reflects the CNCI distribution for papers authored by researchers with a high level of influence. In contrast, as shown in Figure 10(b), when setting aus_factors = C2, the probability distribution of CNCI is 23.10%, 26.50%, 26.10%, and 24.30%, representing the CNCI distribution for papers authored by researcher combinations with papers of relatively high average impact. These two distributions have different meanings, and naturally, they result in different CNCI probability distributions, even for the same Aminer dataset.



(a) the distribution of CNCI by setting tcF=*high* and tc_max=*high*

(a) the distribution of CNCI by setting *aus_factors*=C2

Figure 10. The distribution of CNCI.

Then, we observe the distributions of these three hidden variables. As shown in Figure 11, in the Aminer paper dataset, the independent distribution of *C*1 *aus_factors* is 25.90%, the independent distribution of *C*1 *inst_factors* is 25.90%, and the independent distribution of *C*1 *paper_factors* is 24.1%. The independent distribution of *C*4 *aus_factors* is 30.6%, the independent distribution of *C*4 *inst_factors* is 23.30%, and the independent distribution of *C*4 *paper_factors* is 23.9%.

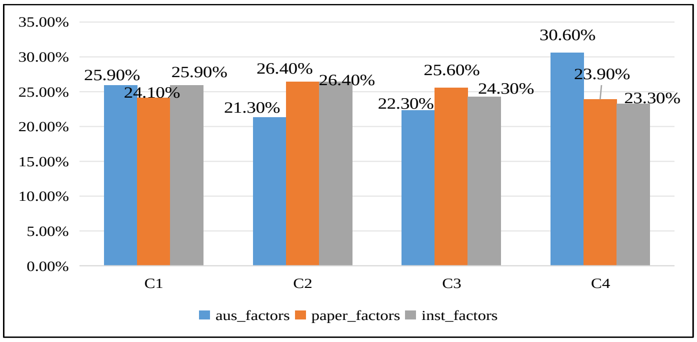
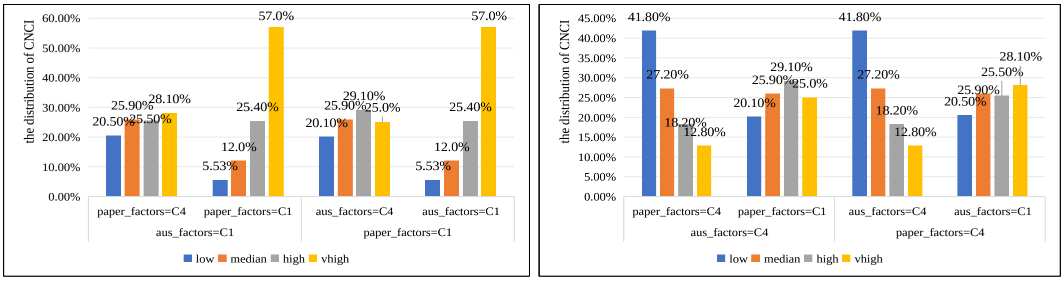


Figure 11. The independent distribution of latent variables.

We also can infer associations of latent variables with *CNCI* and the associations among themselves. Firstly, we first analyze which of *aus_factors* and *paper_factors* has a greater impact on *CNCI*. We find that the characteristics of authors on *CNCI* is slightly higher than the impact of internal features within the paper on *CNCI*. As shown in Figure 12(a), when we set *aus_factors*=*C*1 and change *paper_factors* from *C*4 to *C*1, the probability of *vhigh CNCI* increases from 28.10% to 57%, indicating a change of 28.9%. Similarly, when we set *paper_factors*=*C*1 and change *aus_factors* from *C*4 to *C*1, the probability of *vhigh CNCI* increases from 25% to 57%, indicating a change of 32%. This suggests that, compared to internal features within the paper, author characteristics has a slightly higher impact on the paper's influence (*CNCI*). As depicted in Figure12(b), when both *aus_factors* and *paper_factors* are set to *C*4 and the same operations are performed, the same conclusion is reached.



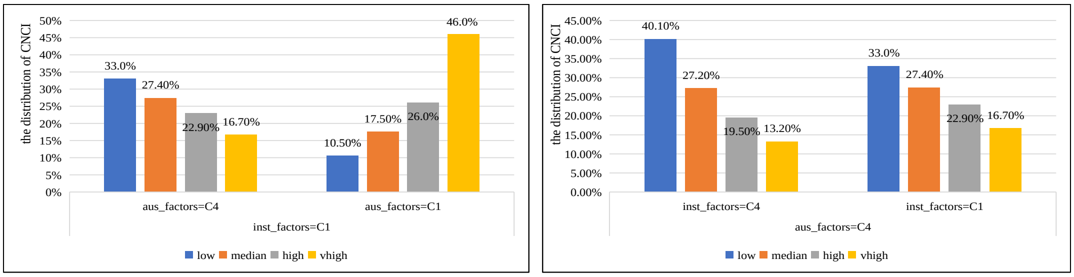
(a)the distribution of CNCI when setting aus_factors=C1 (paper_factors=C1) and changing paper_factors from C4 to C1 (aus_factors from C4 to C1)

(b)the distribution of CNCI when setting aus_factors=C4 (paper_factors=C4) and changing paper_factors from C4 to C1 (aus_factors from C4 to C1)

Figure 12. Distribution of *CNCI* by setting various *aus_factors* and *paper_factors* values.

Next, we analyze the extent to which *aus_factors* and *inst_factors* affect *CNCI*. In addition, we also observed that, compared to institutional characteristics, author characteristics has a greater impact on the paper's influence. Furthermore, through inference, we speculate that within institutions, especially in *C*1 *inst_factors*, the most significant factor in altering the influence of a paper remains the prominence author characteristics within the institution. This underscores the idea that authors, rather than institutions, are fundamentally one of the most influential factors affecting the impact of a paper. In Figure 13(a), when *inst_factors* are fixed at *C*1 and *aus_factors* are changed from *C*4 to *C*1, there is a significant increase in *vhigh*

CNCI (16.7% → 46%). In Figure 13(b), when *aus_factors* are fixed at *C4* and *inst_factors* are changed from *C4* to *C1*, there is only a minor increase in *vhigh CNCI* (13.2% → 16.7%). This suggests that within *inst_factors*, especially in *C1 inst_factors*, *aus_factors* remains the primary factor in altering the impact of a paper.



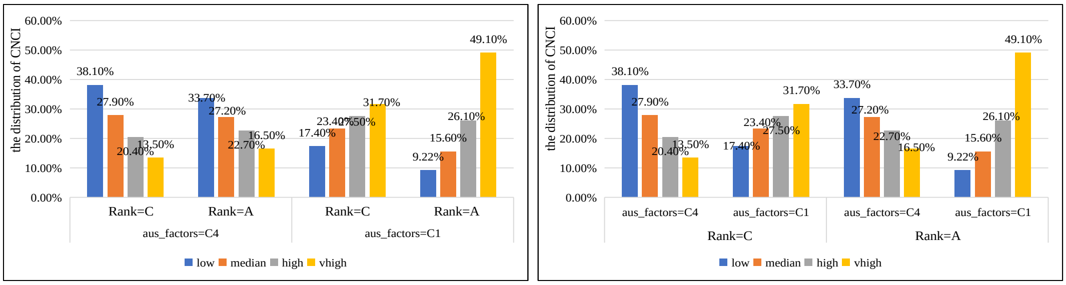
(a) the distribution of CNCI by setting inst_factors=C1 and changing aus_factors=C4 to C1

(b) the distribution of CNCI by setting inst_factors=C4 and changing aus_factors=C4 to C1

Figure 13. Distribution of CNCI by fixing inst_factors and set various aus_factors values.

In addition, many scholars have observed a significant positive correlation between the influence of the publications where papers are published and *CNCI* (Xie et al. 2019; Stegehuis et al., 2015; Didegah and Thelwall 2013). Therefore, we also analyzed the extent to which *aus_factors* and *Rank* affect *CNCI*. We also found that, compared to the *Rank*, author characteristics has a greater impact on *CNCI*. In Figure 14(a), when *aus_factors* are set to *C4* and *Rank* is changed from *C* to *A*, the probability of *vhigh CNCI* increases from 13.5% to 16.5%, with a relatively small increase. In Figure 14(b), when *Rank* is set to *C* and *aus_factors* are changed from *C4* to *C1*, the probability of *vhigh CNCI* increases from 13.5% to 31.7%, indicating a relatively large increase. This suggests that, compared to *Rank*, *aus_factors* have a greater influence on the impact of the paper.

In conclusion, author characteristics is the most critical factor influencing *CNCI*. Compared to the intrinsic features of papers, the influence of author characteristics on *CNCI* is slightly higher. Within institutions, especially in *C1 inst_factors*, the most significant determinant of *CNCI* remains the author characteristics within the institution. Furthermore, in comparison to the *Rank*, the influence of author characteristics on *CNCI* is more pronounced.



(a) the distribution of CNCI by setting aus_factors=*C4*(*C1*) and changing *Rank* from *C* to *A*

(b) the distribution of CNCI by setting *Rank*=*C*(*A*) and changing aus_factors from *C4* to *C1*

Figure 14. Distribution of CNCI when different combinations of aus_factors and inst_factors are set.

Conclusion

In this paper, we adopt a BN with latent variables as the knowledge framework, using latent variables to describe characteristics of different aspects (i.e., institution, author and paper aspects) as a whole, so that interactions among latent category factors as well as observable factors can be analyzed. We use the K-means algorithm to acquire categories of latent variables and use constraint-based scoring approach to learn the BN. We analyzed how the introduction of latent variables provides new perspectives compared to using only observable variables, and conducted corresponding analyses, resulting in certain conclusions.

Leveraging BN with latent variables for inference has allowed us to derive the similar conclusions presented in Sun et al., (2023). However, the inclusion of latent variables has yielded more insights. For instance, within certain institutions, even in *C*1 *inst_factors*, author characteristics remain the primary factor influencing the impact of a paper. Compared with conclusion that authors have greater influence than institutions in Sun et al., (2023), our findings provide a deeper understanding of the interaction between institutions, authors, and CNCI. Additionally, we have uncovered some novel insights, such as from the perspective of papers, author characteristics are the key factors influencing CNCI, surpassing institutional features and paper content.

The data used for the BN construction comes from the Aminer dataset, implying that the research results of this paper are generally applicable to the field of computer science. Exploring the different models or pathways in different fields would be worthwhile in the future. Although this paper comprehensively uses latent variables to represent institutional factors, author factors, and internal paper features, concepts such as institutional influence, scholarly achievements, and paper innovation are complex. We only utilize a portion of bibliometric indicators to represent them, which may result in an incomplete understanding of domain knowledge.

In this study, we ignore the impact of time factor on citations and their categories. However, the temporal factor is crucial to understanding the interactive relationships between paper citations and their category influencing factors . In subsequent research, we will need a framework to study the dynamic interactive relationships between paper citations and their category influencing factors.

Acknowledgments

The work is supported by the Literature and Information Capacity Building Project of Chinese Academy of Science (No. E3291106).

References

Abramo, G., D’Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of Informetrics, 13(1), 32-49.

Amjad T, Shahid N, Daud A, et al. Citation burst prediction in a bibliometric network[J]. Scientometrics, 2022, 127(5): 2773-2790.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. American sociological review, 829-846.

Bornmann, L. (2011). Scientific peer review. Annual review of information science and technology, 45(1), 197-245.

Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. Journal of Informetrics, 9(3), 419-429.

Boyd, B. K., Finkelstein, S., & Gove, S. (2005). How advanced is the strategy paradigm? The role of particularism and universalism in shaping research outcomes. Strategic Management Journal, 26(9), 841-854.

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. Quantitative Science Studies, 2(1), 155-183.

Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov), 507-554.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1), 3741-3782.

Daly, R., Shen, Q., & Aitken, S. (2011). Learning Bayesian networks: approaches and issues. The knowledge engineering review, 26(2), 99-157.

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of informetrics, 7(4), 861-873.

Elidan, G., & Friedman, N. (2013). Learning the dimensionality of hidden variables. arXiv preprint arXiv:1301.2269.

Elidan, G., Lotner, N., Friedman, N., & Koller, D. (2000). Discovering hidden variables: A structure-based approach. Advances in Neural Information Processing Systems, 13.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., & Sebag, M. (2018). Learning functional causal models with generative neural networks. Explainable and interpretable models in computer vision and machine learning, 39-80.

He, C., Yue, K., Wu, H., & Liu, W. (2014, November). Structure learning of bayesian network with latent variables by weight-induced refinement. In Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning (pp. 37-44).

Hurley, L. A., Ogier, A. L., & Torvik, V. I. (2013). Deconstructing the collaborative impact: Article and author characteristics that influence citation count. Proceedings of the American Society for Information Science and Technology, 50(1), 1-10.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—article, author, or journal?. Academy of management journal, 50(3), 491-506.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—article, author, or journal?. Academy of management journal, 50(3), 491-506.

Kalisch, M., & Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research, 8(3).

Kan, Y., Yue, K., Wu, H., Fu, X., & Sun, Z. (2022). Online learning of parameters for modeling user preference based on bayesian network. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 30(02), 285-310.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

Kong, H., & Wang, L. (2023). Flexible model weighting for one-dependence estimators based on point-wise independence analysis. Pattern Recognition, 139, 109473.

Kulis, B., & Jordan, M. I. (2011). Revisiting k-means: New algorithms via Bayesian nonparametrics. arXiv preprint arXiv:1111.0352.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Harvard university press.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers?. Trends in ecology & evolution, 20(1), 28-32.

McCabe, M. J., & Snyder, C. M. (2014). Identifying the effect of open access on citations using a panel of science journals. Economic inquiry, 52(4), 1284-1300.

Mingers, J., & Xu, F. (2010). The drivers of citations in management science journals. European Journal of Operational Research, 205(2), 422-430.

Moed, H. F., & Garfield, E. (2004). In basic science the percentage of "authoritative" references decreases as bibliographies become shorter. Scientometrics, 60, 295-303.

Mourad, R., Sinoquet, C., Zhang, N. L., Liu, T., & Leray, P. (2013). A survey on latent tree models and applications. Journal of Artificial Intelligence Research, 47, 157-203.

Podsakoff, P. M., MacKenzie, S. B., Bachrach, D. G., & Podsakoff, N. P. (2005). The influence of management journals in the 1980s and 1990s. Strategic management journal, 26(5), 473-488.

Qi, Z., Yue, K., Duan, L., Hu, K., & Liang, Z. (2022). Dynamic embeddings for efficient parameter learning of Bayesian network with multiple latent variables. Information Sciences, 590, 198-216.

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International journal of data science and analytics, 3, 121-129.

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?. Scientometrics, 94(1), 57-73.

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?. Scientometrics, 94(1), 57-73.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. Journal of Informetrics, 14(3), 101039.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Shao, Z., Zhao, R., Yuan, S., Ding, M., & Wang, Y. (2022). Tracing the evolution of AI in the past decade and forecasting the emerging trends. Expert Systems with Applications, 209, 118221.

Song, K., Yue, K., Wu, X., & Hao, J. (2021). An efficient approach for parameters learning of bayesian network with multiple latent variables using neural networks and p-em. In Collaborative Computing: Networking, Applications and Worksharing: 16th EAI International Conference, CollaborateCom 2020, Shanghai, China, October 16–18, 2020, Proceedings, Part I 16 (pp. 357-372). Springer International Publishing.

Song, Y., Situ, F., Zhu, H., & Lei, J. (2018). To be the Prince to wake up Sleeping Beauty: The rediscovery of the delayed recognition studies. Scientometrics, 117, 9-24.

Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. Journal of informetrics, 9(3), 642-657.

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. International Journal of Research in Marketing, 32(1), 64-77.

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. International Journal of Research in Marketing, 32(1), 64-77.

Sun, M., Ma, T., Zhou, L., & Yue, M. (2023). Analysis of the relationships among paper citation and its influencing factors: a Bayesian network-based approach. Scientometrics, 128(5), 3017-3033.

Sun, M., Yue, M., & Ma, T. (2023). Differences between journal and conference in computer science: a bibliometric view based on Bayesian network. Journal of Data and Information Science, 8(3), 47-60.

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. Journal of informetrics, 12(1), 203-216.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008, August). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990-998).

Wang, F., Fan, Y., Zeng, A., & Di, Z. (2019). Can we predict ESI highly cited publications?. Scientometrics, 118, 109-125.

Wang, M., Wang, Z., & Chen, G. (2019). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. Scientometrics, 119, 1575-1595.

Wang, Y., Zhang, N. L., & Chen, T. (2008). Latent tree models and approximate inference in Bayesian networks. Journal of Artificial Intelligence Research, 32, 879-900.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378-382.

Wu Xinran & Yue Kun. (2023). Bayesian network learning method with latent variables: Research review. Journal of Yunnan University (Natural Science Edition) (02), 298-313.

Xie, J., Gong, K., Li, J., Ke, Q., Kang, H., & Cheng, Y. (2019). A probe into 66 factors which are possibly associated with the number of citations an article received. Scientometrics, 119, 1429-1454.

Yue, K., Wu, X., Duan, L., Qiao, S., & Wu, H. (2020). A parallel and constraint induced approach to modeling user preference from rating data. Knowledge-Based Systems, 204, 106206.

Zhang, L., & Guo, H. (2006). Introduction to Bayesian Networks. Science Press.

Zhu, S., Ng, I., & Chen, Z. (2019). Causal discovery with reinforcement learning. arXiv preprint arXiv:1906.04477.