Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Datateknik
2024 | LIU-IDA/LITH-EX-A--24/001--SE

Comparison of Causal Models for Bibliometric and Scientometric Analysis Applications

–

Jämförelse av orsakssambandsmodeller för bibliometriska och scientometriska analysapplikationer

–

Hirbod Gholamniaetakhsami

Supervisor : Krzysztof Bartoszek
Examiner : Anders Nordgaard

External supervisor : Kostiantyn Kucher

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

Keyword analysis in scientific articles is a method used to identify and evaluate the importance and relevance of specific words or phrases (keywords) within scientific literature. The primary goal of keyword analysis is to uncover the core themes, research trends, and conceptual frameworks within a given field or across multiple disciplines. It helps researchers understand scientific discourse's focus and ideas' evolution over time.
This thesis performs keyword analysis on a repository of scientific publications through a combination of methods. It starts with extracting the available keywords, and it deals with the missing keywords data through data augmentation. Then, it utilizes a variety of statistical methods to gain insight into the publications.
The study employs an implementation of LDA topic modeling to accurately categorize keywords into thematic groups, a Vector autoregression to explore keyword relationships, and temporal dynamics of keywords. Next, the research further examines the interdisciplinary connectivity of keywords, clarifying the collective nature of modern science.

In conclusion, the thesis presents a comprehensive framework for keyword analysis in scientific literature, Through a blend of data augmentation, natural language processing, temporal dynamics, and interdisciplinary examination, the study provides a robust tool for understanding the development and structure of scientific literature. The findings of this research have important implications for scholars, it allows navigating the vast amount of scientific literature more effectively and to discern the most influential ideas and trends shaping target fields. The methodologies implemented here offer an opportunity for any studies to methodologically search, extract, and identify keywords to find relevant papers and interpret the complex landscape of scientific communication.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

In the rapidly evolving landscape of scientific research, the ability to identify and understand emerging trends is of paramount importance. This is particularly true in the realm of bibliometrics and scientometrics, where the sheer volume of published works can make discerning these trends a daunting task.
Bibliometrics and scientometrics are quantitative methods used to analyze and measure the bulk of scientific publications. They involve statistical analysis of books, articles, or other publications, to study or measure various aspects such as authorship, publication output, or citations.
In the context of bibliometrics, applying quantitative techniques to the relevant metrics (e.g. citations, publication counts, authorship, keywords). The term 'bibliometric' was first used by Alan Pritchard [30]. Through the analysis of bibliometric data, various trends and impacts of specific metrics are expected to be revealed. During the early stages, the well-known bibliometric studies of Lotka's Law, Bradford's Law, and Zipf's Law provided methods for describing the frequency of publication by authors in any given field, estimation of exponentially diminishing returns of searching for references in scientific journals, and estimation of the frequency of a word in a text [92]. The primary goal of bibliometric analysis is to gain insights into the characteristics of publications, which can reveal citation patterns, publication trends, and networks of authorship. Therefore, three examples of what a researcher can achieve by conducting bibliometric analyses are as follows:

1. Identifying research trends: Bibliometric analysis helps identify emerging research trends and areas of interest within a particular field by analyzing the frequency of keywords, topics, or themes in published literature.

2. Assess research Impact: Scholars can evaluate the impact of published articles, journals, or researcher groups by examining citation metrics such as the number of citations received and citation networks.

3. Facilitate literature reviews: Such analyses can help in literature reviews by providing a structured approach to identifying relevant articles and assessing the state of the art within a given field.

Influence and trend analysis are powerful tools used across various fields to understand the evolution and impact of phenomena over time. They provide a dynamic view of a system, revealing how certain elements shape its progression and predicting future developments.
Trend analysis is a powerful tool used across various fields to understand the evolution of a phenomenon over time. In finance, it's used to analyze key financial metrics to understand operational efficiency and firm dynamics [10]. Marketing uses trend analysis to study changes in consumer behavior and preferences [94]. In public health, it's used to track the incidence of diseases over time, helping to identify major health risks and guide future public health policies [77]. In climate science, trend analysis helps in studying changes in weather patterns and temperatures over time, aiding in understanding the impacts of climate change and guiding mitigation strategies [72]. Social media utilizes trend analysis to understand the popularity of certain topics or the sentiment towards certain products or events [105]. In each of these fields, trend analysis provides valuable insights that can guide decision-making and strategy development. This involves the systematic tracking of changes and patterns, providing valuable insights into past behaviors and potential future developments. On the other hand, influence analysis is a powerful tool used across various fields to understand the impact of certain factors on a system or process. It involves identifying key elements or variables and assessing how changes in these elements ripple through the system. In business, for example, influence analysis might be used to understand the impact of demographic variables on small start-ups [75]. In political economy, it could be used to assess how legislative changes might impact public policy [81].
In the context of scientific research, trend analysis can help identify emerging areas of interest, track the progression of specific topics, and predict future research directions [127]. In the realm of scientific research, influence analysis takes on a unique role. In this thesis, it is used to understand how certain themes or topics have shaped the research landscape [102, 89].
Influence analysis and trend analysis play crucial roles in bibliometric studies. Influence analysis helps us understand how certain themes or topics have shaped the research landscape. Influence analysis can reveal the most impactful research works, authors, or institutions in a specific field. On the other hand, trend analysis can help identify emerging areas of interest, track the progression of specific topics, and predict future research directions. Both of these analyses provide valuable insights that can guide decision-making and strategy development in research policies and practices. Keyword analysis can highlight differences rather than similarities, and examining dispersion patterns, concordances, and key clusters can help overcome some limitations of scientometric studies [7]. Also, keywords serve as a concise summary of a paper's content and reflect the main themes of the research; they can reveal new bibliometric indicators and approaches that can help understand discipline development stages [113]. In scientometric research such as [90], keywords form a crucial part of the metadata that accompanies each scientific publication and facilitate the analysis.

1.2 Aim

Scientometrics, the study of scientific publications and their impact, relies heavily on keyword analysis to uncover trends, emergent themes, and influential factors shaping the research landscape. This project aims to look into the dynamics of keywords within scientific literature using statistical methods, particularly focusing on causality analysis techniques. In this research, we draw inspiration from the complex applications of trend and influence analysis across various domains. This research tries to adapt these methodologies to the realm of scientometrics. By systematically tracking changes and patterns in keywords, valuable insights into past research behaviors and potential future developments can be revealed. To summarize, the two most important objectives of this project are:

1. **Trend and emergence analysis of keywords**: This objective focuses on the evolution and emergence of keywords within the scientific literature. By employing statistical

techniques, you aim to identify patterns of keyword usage, detect emerging areas of interest, and predict future research directions.

2. **Influence analysis through causality methods**: This objective aims to understand the influence of keywords on each other within scientific publications. Utilizing causality analysis, we aim to identify key elements or variables and assess how changes in these keywords affect the overall research landscape.

1.3 Research questions

This research is designed to tackle two primary points that hold significance both from a pragmatic viewpoint and a scholarly standpoint. The following questions aim to look into the dynamics of keywords within the scientific literature, gaining insights that can guide decision-making in research policies and practices:

1. **Trend and Emergence Analysis of Keywords:**
    - What are the most frequently used keywords in scientific literature over the past decades, and how has their usage changed over time?
    - Can we identify the themes that have emerged in the last five years from keywords, and what keywords are associated with these themes?
2. **Influence Analysis through Causality Methods:**
    - How does a surge in the usage of a particular keyword influence the usage of other related keywords in scientific publications?
    - Is there a general pattern for causal relationships between keywords?

1.4 Delimitations

This thesis is focused on the analysis of the dataset of papers from researchers at Linköping University, which is part of the DiVA institutional repository. DiVA is a repository for research publications and student theses written at 50 universities and research institutions. However, this study will only consider the documents from Linköping University for the following reasons:

- **Focused Analysis:** By limiting the dataset to Linköping University, the research can provide a more focused and detailed analysis. This allows for a deeper understanding of the trends, emergence, and influence of keywords within a specific institutional context.
- **Manageable Scope:** The DiVA repository contains documents from around 50 institutions, which could make the scope of the project too broad and time-consuming for a detailed analysis. By focusing on one university, the project remains manageable for more precise conclusions within the timeframe.
- **Relevance:** We hypothesize that Linköping University may have specific areas of research focus or institutional practices that influence keyword usage. Analyzing this specific dataset can provide insights relevant to Linköping University and similar institutions.

2 Literature review

In essence, Scientometrics strives to measure the evolution of scientific domains, the impact of corresponding publications, and the patterns of authorship.
It's important to note that in practice, Scientometrics shares a significant overlap with Bibliometrics.
There have been many previous studies on bibliometrics and scientometrics.
[46] provided insight into the citation analysis during the early years of the Journal of Consumer Research. During their study, authors analyzed the most frequently published researchers(with four or more papers), and developed a model comprising two stages(Asymmetry correction and citation-similarity space) to investigate the patterns. The results showed two substantial differences in the levels of analysis(individual vs social) and the methodology of research(experimental-oriented vs mathematical-oriented).
In regards to Domain analysis, [115] provided an analysis of frequently cited authors for the period of 1972-1995 in an attempt to visualize the domain of information science. Through 3 statistical routines of (1)Factor Analysis (2)Multidimensional Scaling and (3)Cluster Analysis, the authors identified two major subdisciplines of information science: bibliometrics (including citation analysis) and retrieval (including user studies and system design), which showed little integration or overlap over the 24 years.
In [53], Hullmann and Meyer studied the field of nanotechnology through bibliometric analysis and scientometrics. They utilized publications and patents as indicators, investigating their evolution and dispersion across various countries, disciplines, and sectors. The research process involved conducting keyword searches in relevant databases and subsequently employing a classification scheme to allocate publications and patents to distinct scientific fields. The findings of the study underscored that nanoscience and nanotechnology are not only highly multidisciplinary but also rapidly expanding fields. Moreover, the patterns of publication and patenting exhibited considerable variation among different regions and organizations. In their study, referenced as [56], Jiancheng and Junxia adopted a different approach to assessing the performance of research groups within the domain of information science in China. They introduced a DEA analysis model, using the budget and size of the research groups as input parameters. The output was determined by the quantity and quality of publications, gauged through six distinct indicators. The authors posited that enhancing the efficiency of knowledge production necessitates the selection of knowledgeable workers, effective utilization of Information Technology, and the cultivation of a conducive knowl-

edge culture. [39] investigates the relationship between the ranking of universities and the number of citations belonging to their leaders. According to this article, there is a positive correlation between the research citation of leaders of the top hundred universities and the ranking. It is noteworthy to say that, US university presidents tend to have higher citation levels than presidents of other universities in this article. In the research referenced as [20], the authors analyze various aspects of tsunami research such as language, document type, publication output, authorship, publication patterns, subject category, author keywords, country of publication, and citation impact. As a result, it was shown that the dominant language of articles is English, Pure and Applied Geophysics is the most active journal, and Multidisciplinary Geosciences is the most common subject category for tsunami research.
Apart from the abovementioned studies, some studies specifically present theories or practical approaches to facilitate the process of bibliometrics. [13] reviews visualization techniques for mapping the domain structure of scientific disciplines and the bibliographic structure of the field itself. In addition, the authors provide details about the general process flow of visualizing knowledge domains, which involves data extraction, unit of analysis, measure selection, similarity calculation, ordination, and visualization. [110] discusses the advantages and limitations of using bibliometric methods. Wallin reviews different types of publication pattern analysis, such as bibliographic, journal impact factor, and publication types, and how they can reflect the quality and visibility of research output. The author suggests that bibliometric indicators should be complemented by other methods, such as peer review, expert judgment, text mining, and patent analysis, to provide a more comprehensive and balanced assessment of research. [18] aims to develop an approach to detect and visualize emerging trends and transient patterns in scientific articles. The resulting visualization system-CiteSpace II - can reveal the evolution of research fronts, the impact of internal and external events, the identification of pivotal points, and the verification of domain experts. [68] studies the identification of emergence for research fields using the scientometric analysis. The methodology involved applying the co-word analysis technique and creating a knowledge map that reveals the patterns, then applying the network analysis to measure the centrality and finding active research fields. This method proved to be successful according to the case study done for the information security field, in which case some identified potential hubs for future research are secure communication, digital signature, and cryptosystem. [107] proposes a general text mining approach for patent analysis. A general strategy for patent analysis follows the following steps:

- Data collection: After defining the scope of analysis, search and acquire relevant patents from various databases.
- Data prepossessing: segment and summarize the documents. Extract keywords, phrases, and other features for indexing and analysis.
- Data identification: Apply various measures to group the patents and find out the underlying concepts and categories.
- Data interpretation: Analyze the results to discover patterns, insights, and intelligence for decision-making.

In this research, the methodology suggested by the authors includes defining the task, searching, filtering, and downloading relevant patents, segmenting, abstracting, and clustering the patent content, and creating and interpreting visualized results, such as patent maps. The paper also evaluates some of the critical techniques. It demonstrates their feasibility on a real-world patent set for domain analysis and mapping, which shows that their approach is more effective than existing classification systems.
In the book [78] the theory and applications of temporal data mining in various fields are explored. Such a data mining approach can be used in bibliometrics to analyze the development

of a field over time. In [101] authors used temporal analysis to gain insight into the evolution of the field, They also used visualization multidimensional scaling, and parallel coordinate analysis in conjunction with temporal analysis to reveal the development phases of institutional repositories. [36] introduces two models: the Influence Model (DIM) and the Dynamic Topic Model (DTM) to analyze the influence of documents over time and understand how language evolves in a corpus. Using the proposed methods it is possible to measure the impact of a document through its influence on future works. Using the probabilistic influence model the authors showed that the posterior influence scores of the model are significantly correlated with citation counts, it was also revealed qualitatively different aspects of influence that citation counts may miss. [119] is a study that explores the cosine similarity among six types of scholarly networks. According to the input data containing 59 journals from 1965 to 210, the topical networks and co-authorship networks have the lowest similarity, while citation networks and co-citation networks have high similarity. [121] applied four topic modeling algorithms(LDA, CTM, Hierarchical LDA, and HDP) on a test sample comprising seven different scientific areas. To evaluate the clustering performance, authors used precision, recall, and F score, and compared the results with a baseline method constructed by the K-means algorithm. In the end, they concluded that HDP was able to infer the optimal number of topics and achieve high accuracy and precision, and therefore performs best in the targeted domains. [66] presented local and global maps of career topics based on the term co-occurrences in the abstracts and titles of 3141 articles on careers published in management journals and 16,146 articles on careers published in social science journals between 1990 and 2012. Each map showed six clusters of career topics across relevant fields. This study provides the potential for scholars to explore the literature, find research opportunities, and exchange insights across different sub-fields and disciplines.
In [123], the authors construct a feature space of 24 variables for 1,025 papers in Information Science & Library Science published in 2007. They apply stepwise multiple regression analysis and ten-fold cross-validation to select the optimal model for explaining the relationship between the features and the citation impact after 5 years of publication. The paper argues that objective features of papers can make citation impact prediction relatively accurate. Finally, the authors acknowledge the limitations and caveats of their study, such as the sample size, the data source, and the omitted features. The study by Bornmann (2015) [14] provides a comprehensive review of research on three distinct types of altmetrics: microblogging (specifically Twitter), online reference managers (namely Mendeley and CiteULike), and blogging. The paper employs a meta-analysis approach to determine the correlation between the counts of altmetrics and traditional citations for each type of altmetric. The findings reveal that the correlation is virtually non-existent for Twitter, relatively minor for blogging, and ranges from moderate to substantial for online reference managers. [41] discusses the Taverna workbench, an open-source scientific workflow manager, that can be used to integrate various tools and data sources for bibliometric analysis, such as Web services, XML parsers, R packages, and visualization tools. Additionally, the authors share their workflows on the myExperiment platform, enabling the reusability and reproducibility of bibliometric analyses. [103] proposes an investigation on the quality and coherence of generated topics depending on the type of textual data(abstract or full-text). The datasets under study contain scientific publications from the domain of fisheries, one with 4,417 articles from a single journal and another with 15,004 articles from 12 journals. The paper finds that full-text data produces more coherent and interpretable topics than abstract data, especially for smaller datasets. The paper also finds that document frequency, document word length, and vocabulary size have mixed effects on topic coherence and human topic ranking. [55] is centered on the application of Natural Language Processing (NLP) to enhance citation analysis. The authors created annotated datasets for four tasks, namely citation purpose and polarity, citation context, reference scope, and sentiment analysis. The results of their research have potential applications in various scientometric tasks. They demonstrated how these results can be useful for applications such as scientific summarization, measuring re-

search dynamics, and survey generation. Furthermore, they showed how the annotations can help generate more focused and informative summaries of scientific papers and fields. [59] propose three applications that combine text mining and foresight methods, such as road mapping, scenario development, and comparing public and scientific discourse, to detect and examine emerging topics and technologies, and to provide a solid base for decision-making. This study suggests that Text mining can complement and enhance the qualitative and participatory character of foresight, by integrating more views and stakeholder positions. [80] bibliometric study to analyze the intellectual structure of the smart city research field from 1992 to 2012. Authors use two hybrid techniques that combine co-citation clustering and text-based analysis to identify and label thematic clusters of publications, As a result, they identified five development paths supported by the thematic cluster of publications. Furthermore, by analyzing the strategic principles of each path, they find four terms of divisions(dichotomies). [3] proposes a novel neural network model that learns the citation patterns from a large dataset of over 1.6 million papers from different research fields and their citations. The authors show that the proposed sequence-to-sequence model outperforms the baseline methods (MEY, AVR, and GMM) in terms of prediction accuracy, correlation, and error metrics. [50] aims to identify themes and compare differences in online travel reviews from three major OTAs in China using semantic association analysis. The authors use natural language processing, statistical analysis, and social network analysis to extract thematic words, construct bigrams of co-occurrence phrases, and measure the structural properties of semantic association networks. They conclude that there are apparent discrepancies among the three platforms in terms of thematic words, topic distribution, network density, modularity, and small-world properties. [2] propose a methodology to predict the long-term impact of a scientific article soon after its publication, using a combination of early citations and journal impact factor. Two linear regression models, one based on rescaled citation counts and another on log-transformed citation counts, are used to test the hypothesis. The authors suggest that the prediction accuracy is good for citation time windows above two years, decreases for lowly-cited publications, and varies across disciplines. [62] compares different methods for generating publication embeddings and their resulting relatedness measures with conventional citation-based and abstract-based publication relatedness measures. The paper evaluates four embedding-derived relatedness measures, based on word2vec embeddings of citation labels, sentence embeddings using BERT and SciBERT, and title and abstract embeddings using SPECTER, and compares them with two conventional relatedness measures, one based on BM25 text similarity of title and abstract noun phrases, and one based on combined direct citation, bibliographic coupling and co-citation (DC-BC-CC). Authors find that embedding-derived relatedness measures resemble citation-based relatedness measures more strongly than text-based relatedness measures and that they also outperform conventional techniques in clustering publications cited with the same citation intent. [71] proposes a novel method for detecting research trends by predicting the frequency of author-defined keywords (AKs) in scientific publications. The proposed author-defined keyword frequency prediction task (AKFP), is based on the long short-term memory (LSTM) neural network and four categories of features: temporal feature, persistence, community size, and community development potential. Evaluating the model's performance against the baselines reveals that, the AKFP achieves better accuracy and generalization than the baselines, especially for long-term prediction. [5] proposes a novel approach to identify important citations using sentiment analysis of in-text citations and cosine similarity of paper contents. The article concludes that the proposed in-text citation sentiment analysis technique can identify the important and non-important citations with more accuracy than the state-of-the-art techniques and that it can be useful for various applications such as citation recommendation. [26] provides guidelines for conducting bibliometric analysis, which include defining the aims and scope of the study, selecting the appropriate techniques and data sources, collecting and cleaning the data, running the analysis, and reporting the findings. Additionally, they present a toolbox of available techniques for bibliometric analysis, next they explain the usage, unit of analy-

sis, and data requirements of each technique, such as citation analysis, co-citation analysis, bibliographic coupling, co-word analysis, and co-authorship analysis. [49] proposes a new model, called MDER, which combines rule embedding and a CNN+BiLSTM-Attention-CRF structure to effectively extract the method and dataset entities from the main textual content of scientific papers. The authors construct four new datasets from four research areas of computer science, namely NLP, computer vision, data mining, and artificial intelligence, and evaluate the proposed model on these datasets. It is shown that the proposed model outperforms the state-of-the-art approaches and has good transferability and generalization performance across different areas. [1] proposes a novel framework that can rank the nodes (papers) in temporal citation networks by learning low-dimensional node embeddings and a probabilistic regression model. The framework can handle dynamic graphs, temporal features, and ranking problems. The model was evaluated on an arXiv paper citation network using six standard information retrieval-based metrics, such as AUC, Kendall's rank correlation, precision, novelty, temporal novelty, and NDCG. The results showed that the proposed model outperforms baseline Support Vector Regression (SVR), Time-Balanced PageRank (TBP), and PageRank (PR) on average.
[69] proposes a method that parses the syntactic dependency relations between words to extract research topics from 4,182 abstracts of accounting articles published from 2000 to 2019. Also, they identified cold and hot topics through an AR(1) autoregressive model. The authors suggest that results are similar to those of LDA topic modeling but more efficient and interpretable.
Scientific publication data is a rich source of information that can reveal various aspects of the research landscape, such as trends, impact, collaboration, and innovation. In the information visualization and visual analytic research communities, such publication data analysis problems are relevant from at least two perspectives: (1) designing and applying (interactive) visualization approaches to gain insights from the data [31], and (2) conducting analyses on the publication data *on* the respective research communities.
The first perspective involves techniques, such as sentiment visualization [61] and data embedding methods [52].
Sentiment visualization allows for the graphical representation of sentiment data, providing a clear picture of the emotional tone within the data. Data embedding methods, on the other hand, are used to reduce the dimensionality of the data, making it easier to visualize and analyze. The second perspective, as discussed in the abovementioned topics involves studying publication trends, author collaborations, citation networks, and other bibliometric indicators within these communities.
A common approach to analyzing scientific publication data is to use correlation analysis and pattern mining. Correlation analysis in scientific articles is a statistical method used to evaluate the strength of the relationship between two quantitative variables [54]. Through this method, it is possible to gain insights into the correlational relationship between variables and the strength of those relationships. On the other hand, pattern mining is a method to discover interesting and relevant patterns in data [35], hence it can be used to uncover various kinds of knowledge from diverse data types and it is known for effectiveness.
In the context of scientific articles, correlation analysis could be used to determine if there's a relationship between the number of times an article is cited and the impact factor of the journal it's published in. On another note, pattern mining could be used to discover interesting associations in the data related to scientific publications, for instance, it could help to identify if certain keywords tend to appear together in articles, or if certain topics have a rise in popularity over time.
While both methods are very useful in gaining insights into hidden knowledge in a given data, they fall short in explaining the 'why' behind the observed relationship. This is where causality analysis comes in.
Causal analysis goes a step further by trying to understand the cause-and-effect relationships

between variables [88]. Often three key assumptions for causality are considered in the literature [120].

- Stable Unit Treatment Value Assumption (SUTVA): The first is that the treatment of one unit does not affect the outcome of another unit.
- Ignorability (or Unconfoundedness): This assumption states that the assignment to a treatment is independent of the potential outcomes.
- Positivity (or Overlap): This assumption requires that each unit has a positive probability of receiving the treatment, given the covariates.

Therefore Causality modeling can complement the existing correlation analysis and pattern mining approaches that do not directly take the temporal or causal aspects into account.

3 Theory

The purpose of this Chapter is to give an overview of the methodology for performing the two approaches of trend investigation and influence analysis. The first section 3.1 begins the introduction for topic modeling, the primary method to discover the themes from the keywords, the section (3.2,3.3,3.4) would focus on giving an overview of background for the three main methodologies used in this thesis project. Note that this chapter focuses on the basic concepts of these methods and the explanation for the process and the actual implementation will be further explored in the chapter (4,5)

3.1 Topic Modeling

Topic modeling belongs to the category of statistical modeling used in Natural Language Processing. The idea behind these methods is to use the probability distribution over words [83]. Derived from the core idea, there are several topic modeling algorithms available in the literature such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA).

3.1.1 Latent Semantic Analysis

LSA is a technique that helps in understanding and extracting meaning from text by statistically analyzing patterns of word usage across a collection of documents[29].
This method works by performing singular value decomposition (SVD) to a constructed term-document matrix [63] and selecting a lower-dimensional approximation of the text corpus by keeping only the top largest singular values.

3.1.1.1 Mathematical Representation

The key idea behind the three steps of this technique is to select the top k dimension and capture the most significant semantic structure and thus reduce noise. The steps are as follows:

1. **Construct a term-document matrix:** Considering the construct a matrix A of size $m \times n$, we have have $A = \left[a_{ij}\right]$ where:
    - m is a set of unique terms and n documents

- a_{ij} represents the weight of term i in document j and therefore : Term Frequency (TF): $a_{ij} = TF_{ij}$

2. **Apply SVD to decompose A:** The decomposition step generates three matrices to reveal the latent semantic structures:

$$A = U\Sigma V^T$$

where:

- U represents term eigenvectors.
- Σ represents singular values.
- V^T represents document vectors of V

3. **Keeping the top dimensions:** Keeping the k most important dimensions associated with the top k highest singular values, forms a truncated matrix as:

$$A_k = U_k \Sigma_k V_k^T$$

3.1.2 Probabilistic Latent Semantic Analysis

PLSA is an extension of the LSA method where the latent topics are modeled through probabilistic modeling. This method is shown to have higher performance gain than the LSA method [48]. In this model, the conditional probability between documents and words is modeled using latent variables [15].

3.1.2.1 Mathematical Representation

To introduce this concept we need to define certain assumptions:

1. Z is the latent variable, D is the document variable, and W is the word variable.
2. $P(Z \mid W)$: is the probability of a topic given a document.
3. $P(W \mid Z)$ is the probability of a word given a topic. Additionally, this probability is conditionally independent of the document$P(W \mid Z, D) = P(W \mid Z)$).

This technique relies on the estimation of $P(W \mid Z)$ and $P(Z \mid D)$ through the iterative Expectation-Maximization(EM) algorithm with the following steps:

1. **E-step:** compute the posterior probabilities of the latent variable (topic) given the observed data (document and word):

$$P(Z|D,W) = \frac{P(W|Z)P(Z|D)}{\sum_{Z'} P(W|Z')P(Z'|D)}$$

2. **M-step:** update the parameters to maximize the likelihood of the observed data given the posterior distribution:
For words given topics:

$$P(W|Z) = \frac{\sum_D C(D,W)P(Z|D,W)}{\sum_{W'} \sum_D C(D,W')P(Z|D,W')}$$

For topics given document:

$$P(Z|D) = \frac{\sum_W C(D,W)P(Z|D,W)}{\sum_{Z'} \sum_W C(D,W)P(Z'|D,W)}$$

where:

- $P(W,D) = P(D) \sum_Z P(W|Z)P(Z|D)$
- $C(D,W)$ represent the count of word W in document D.

3.1.3 Latent Dirichlet Allocation

Like PSLA, LDA is another generative probabilistic model used in natural language processing and machine learning to discover the underlying topics in a collection of documents. First introduced in [12], LDA also represents topics using word probabilities. In literature, LDA was used in various papers to extract topics from a collection of documents. The papers [42, 86] are examples of works performing this technique in scientific publications. The first paper presents an application of LDA topic modeling aimed to predict research trends based on 3269 articles from the Journal of Applied Intelligence over 30 years and the latter used this method to identify topics from 1000 papers within a corpus of multidisciplinary scientific papers(4 fields), his approach provides a historical perspective as well as forecasts future trends, providing support for the corresponding researchers.

3.1.3.1 Mathematical Representation

As a probabilistic method used in topic model, there is a need to identify the underlying topics in a corpus of documents. To achieve this, LDA needs to explain the observed data (every documents in the corpus) in terms of its latent variables (topics) and the distribution of words for each topic.
To reach this goal, we need to estimate three variables:

1. The topic distribution for each document (θ_d).
2. The word distribution for each topic (β_k).
3. The assignment of topics to words within documents (z_{dn}).

Assuming each document is a distribution of topics, and each topic is a distribution of words, the generative process in LDA will be as follows:

1. Choose the number of topics K.
2. For each document d:
   - **Document-topic distributions** θ_d: Choose a distribution over topics θ_d from a Dirichlet distribution with parameter α.
   $$\theta_d \sim \text{Dir}(\alpha)$$
   - For each word w_n in document d:
     - **Topic assignments** z_{dn}: Choose a topic z_{dn} from the topic distribution θ_d.
     $$z_{dn} \sim \text{Multinomial}(\theta_d)$$
     - **Word generation** w_{dn}: Choose a word w_{dn} from the distribution over words for topic z_{dn}, which is $\beta_{z_{dn}}$.
     $$w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$$

α and β are the hyperparameters of the Dirichlet distributions that control the distributions over topics and words, respectively.
The next step is to find the parameters that maximize the likelihood of the observed data, therefore the aim would be to compute the probability of the corpus given its parameters (α and β) that maximize this probability. To calculate this, we need to first compute the joint probability of the document (partially [32]):

1. Beginning with Dirichlet distribution for θ_d:

$$p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_{d1}^{\alpha_1 - 1} \theta_{d2}^{\alpha_2 - 1} \cdots \theta_{dK}^{\alpha_K - 1}$$

2. Probability of the topic assignments $\mathbf{z}_d$ given θ_d is computed by:

$$p(\mathbf{z}_d \mid \theta_d) = \prod_{n=1}^{N_d} p(z_{dn} \mid \theta_d) = \prod_{n=1}^{N_d} \theta_{d,z_{dn}}$$

   Here, $\theta_{d,z_{dn}}$ is the probability of topic z_{dn} under the topic distribution θ_d.

3. And given $\beta_{z_{dn},w_{dn}}$ is the probability of word w_{dn} under the word distribution for topic z_{dn}, probability of the words $\mathbf{w}_d$ given the topic assignments $\mathbf{z}_d$ and the word distributions β: would be:

$$p(\mathbf{w}_d \mid \mathbf{z}_d, \beta) = \prod_{n=1}^{N_d} p(w_{dn} \mid z_{dn}, \beta) = \prod_{n=1}^{N_d} \beta_{z_{dn},w_{dn}}$$

4. To combine the three probabilities, we need to compute the joint proabability of the variables, and the joint probability of θ_d, $\mathbf{z}_d$, and $\mathbf{w}_d$ is the product of the individual probabilities:

$$p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) = p(\theta_d \mid \alpha) \cdot p(\mathbf{z}_d \mid \theta_d) \cdot p(\mathbf{w}_d \mid \mathbf{z}_d, \beta)$$

   Substituting with the probabilities from the first 3 steps:

$$p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) = \left(\frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_{d1}^{\alpha_1 - 1} \theta_{d2}^{\alpha_2 - 1} \cdots \theta_{dK}^{\alpha_K - 1} \right) \left(\prod_{n=1}^{N_d} \theta_{d,z_{dn}} \right) \left(\prod_{n=1}^{N_d} \beta_{z_{dn},w_{dn}} \right)$$

   And after simplification:

$$p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) = p(\theta_d \mid \alpha) \cdot \prod_{n=1}^{N_d} p(z_{dn} \mid \theta_d) \cdot p(w_{dn} \mid z_{dn}, \beta)$$

This equation expresses the joint probability of the topic distribution θ_d, the topic assignments $\mathbf{z}_d$, and the words $\mathbf{w}_d$ in terms of the Dirichlet prior α and the word distribution β.

1. Now, to find the probability of the observed words, we need to marginalize over the latent variables θ_d and $\mathbf{z}_d$:

$$p(\mathbf{w}_d \mid \alpha, \beta) = \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

2. Given M documents, the probability of the entire corpus $\boldsymbol{D}$ is the product of the probabilities of each document:

$$p(\boldsymbol{D} \mid \alpha, \beta) = \prod_{d=1}^{M} p(\mathbf{w}_d \mid \alpha, \beta)$$

3. By plugging in the expression for $p(\mathbf{w}_d \mid \alpha, \beta)$:

$$p(\boldsymbol{D} \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

The final equation results in the full marginal likelihood of the corpus under the LDA model, which is the final form of the probability of a corpus given the parameters. This process ensures that the LDA identify the topics that best explains the text corpus.

3.2 Vector Autoregression

VAR or Vector Autoregression, introduced by Sims [22] is a statistical technique that captures the linear interconnections between multiple time series. These models are an extension of univariate autoregression and allow for analysis of multivariate time series. In the field of macroeconomics, VAR models are extensively used in macroeconomics to understand complex interactions among economic indicators, with applications including forecasting and estimation of impulse response functions [37] In the field of finance, VAR models help in analyzing interrelationships and making decisions based on the transition matrices of multiple financial variables over time [126] Neuroimaging data analysis employs VAR models to study brain-network dynamics, capturing time-lagged influences among connected brain regions [19] VAR models are widely employed across various disciplines, including economics, finance, neuroimaging, and other domains that deal with time-series data. They are particularly valuable for analyzing dynamic relationships in time series data, forecasting, and structural analysis.

3.2.1 The Structure of VAR

A VAR model describes the evolution of a set of (k) variables (called endogenous variables) over the same sample period (t = 1, ..., T) as a linear function of only their past values. The variables are typically economic indicators such as GDP, inflation, interest rates, etc. The basic form of a VAR model of order p denoted as VAR(p), can be written as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + \varepsilon_t$$

where:

- y_t is a (k x 1) vector of endogenous variables at a time (t),
- c is a (k x 1) vector of constants (intercepts),
- Each (A_i) (for (i = 1, ..., p)) are (k x k) matrices of coefficients to be estimated,
- and ε_t is a (k x 1) vector of error terms that are assumed to be white noise.

The VAR model requires that the time series data must be stationary. The augmented Dickey-Fuller (ADF) test is a statistical test used to determine if a time series is stationary or non-stationary. The ADF test is based on the following null hypothesis:

$$\begin{cases} H_0\textbf{:} & \text{The time series has a unit root} \\ H_1\textbf{:} & \text{The time series does not have a unit root} \end{cases}$$

The outcome of this test is either non-stationarity (the non-stationary condition can not be rejected) or stationarity (the null hypothesis is rejected). Once the stationarity is established, the VAR model is estimated using time series data for the variables of interest. The estimation process involves determining the values of the coefficients A_i that best fit the historical data. This is typically done using methods such as Ordinary Least Squares (OLS).

3.3 Granger Causality

Granger causality (GC) is a statistical concept that assesses causal influence based on the ability to predict one time series from another. It is founded on the principle that a cause precedes its effect and that knowledge of the cause can improve the prediction of the effect. In the context of linear vector autoregression, if a time series X helps to predict future values of a time series Y beyond the information already contained in the past values of Y, then X is said to Granger-cause Y [85]. This concept is particularly relevant in fields such as economics, neuroscience, and any domain where time series data are analyzed [97].

3.3.1 Mathematical Representation

The core idea of Granger causality is to test whether the past values of one variable XX can help predict the future values of another variable YY better than using the past values of YY alone. This is typically done by comparing two vector autoregressive (VAR) models:

1. Restricted model (only using past values of Y):

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t$$

2. Unrestricted model (using past values of both X and Y):

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{p} \gamma_i X_{t-i} + \varepsilon_t$$

Here, the notation is:

- Y_t is the value of variable Y at time t
- X_{t-i} is the value of variable X at time t - i (i.e., lags back in time)
- α is the constant term
- β_i are the coefficients for the past values of Y
- Y_i are the coefficients for the past values of X
- ε_t is the error term

The null hypothesis (H_0) is that X does not Granger-cause Y, which means that all the γ coefficients are zero. This hypothesis is tested against the alternative hypothesis (H_1) that at least one γ is non-zero, using an F-test.

3.4 Bayesian Network Analysis

Bayesian network analysis, also known as Bayesian networks or probabilistic graphical models, is utilized to model the dependencies among various variables in complex systems. They are particularly useful in situations where data may be incomplete or uncertain.
Bayesian network analysis has a wide range of applications across various fields, in particular, these applications are significant:

1. **Dependability, Risk Analysis, and Maintenance:** Bayesian networks are increasingly used in these areas due to their ability to model complex systems, make predictions and diagnostics, compute exact occurrence probabilities of events, and update calculations according to evidence. In [114] it is noted that the increased trend for using this method is due to their advantages over classical methods like Markov Chains, Fault Trees, and Petri Nets.

2. **Multilevel System Reliability:** They extend the use of Bayesian networks to multi-level discrete data, which is particularly useful when system structures are too complex for representation by fault trees. This allows for joint inference about all nodes in the network [118].

3. **Data Mining:** Bayesian networks can encode probabilistic relationships among variables and handle missing data entries. They are used to learn causal relationships, combine prior knowledge with data, and avoid overfitting [45]. These networks are also useful for understanding domains and predicting the consequences of interventions.

4. **Computational Biology:** In biological sciences, Bayesian networks are important for inferring cellular networks and modeling protein signaling pathways [84, 33], classification of biological data such as gene function prediction classification [84, 116], and building Gene regulatory networks(GRN) [17, 125, 122]. In this way, BNs provide a framework for expressing joint probability distributions and inference, combining domain knowledge with data, and learning from incomplete datasets.

These applications demonstrate the versatility and power of Bayesian network analysis in providing solutions to complex problems across different scientific and engineering domains.

3.4.1 Mathematical Representation

3.4.1.1 Conditional Probability and Bayes' Theorem

Conditional probability is the probability of an event given that another event has occurred. If the event of interest is A and event B has already occurred, the conditional probability of A given B is denoted as $(P(A|B))$ and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where $\mathrm{P}(A \cap B)$ is the joint probability of A and B occurring together, and $(P(B))$ is the probability of B occurring.

Bayes' theorem , on the other hand, relates the conditional and marginal probabilities of random variables. It is expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This theory describes how to update the probability of a hypothesis based on evidence.

3.4.1.2 Bayesian Networks

In principles, Bayesian networks are tools for the computation of probabilities, which quantifies uncertainty about events and relationships between events. A Bayesian network represents random variables as nodes in a graph and their conditional dependencies as directed edges between these nodes.

In a Bayesian network, the joint probability distribution is given by the product of the conditional probabilities of each node given its parents: [11]. For a set of variables $X_1, X_2, ..., X_n$, the joint distribution is defined as:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|\text{parents}(X_i))$$

where $parents(X_i)$ are the parent nodes of X_i in the network.
The factorization of the joint probability distribution is a key property of Bayesian networks, as it allows for efficient inference and learning algorithms. To implement a Bayesian Network analysis for scientometric data, the following steps are involved:

1. **Nodes:** The nodes in a Bayesian network represent the variables of interest, which can be discrete or continuous.
2. **Edges:** The directed edges in the network represent the causal or conditional dependencies between the variables. The direction of the edges indicates the direction of the influence or causality.
3. **Conditional Probability Distributions:** Each node in the network is associated with a conditional probability distribution that describes the probability of the variable taking on different values, given the values of its parent nodes.
4. **Inference:** Bayesian network analysis allows for both forward and backward inference. Forward inference is used to predict the probability of certain outcomes given the values of the input variables, while backward inference is used to identify the most likely causes of observed outcomes.
5. **Learning:** Bayesian networks can be learned from data using various algorithms. Learning in Bayesian networks can be divided into two main tasks:
    - **Structure learning:** A structural learning algorithm is employed to learn the structure of the Bayesian network directly from the data.
    - **Parameter learning:** Estimating the conditional probability tables (CPTs) for each node given its parents, typically using maximum likelihood estimation or Bayesian methods.

In the next section, the structure learning through algorithms is explored.

3.4.2 Structure Learning Algorithms in Bayesian Networks

To determine the network structure representing the dependencies among variables a class of algorithms known as structure learning is involved. According to Kitson et al[60], there are three main categories of structure learning algorithms:

1. **Constraint-based Learning**: The first approach uses statistical testing to identify conditional independencies between variables. The process typically involves three steps:
    - **Identification of Conditional Independencies through statistical testing**: The algorithm performs statistical tests (e.g., Chi-square tests) to decide if two variables are conditionally independent given a set of other variables.
    - **Graph Construction**: Based on the identified independencies, a graph is constructed where nodes represent variables and edges represent dependencies.
    - **Orientation of Edges**: Apply rules to orient the edges to form a Directed Acyclic Graph (DAG).

Two Common constraint-based algorithms include the Peter-Clark algorithm and its variants [98, 99, 23] and the Grow-Shrink algorithm [73]. These methods rely on the assumption that if variables are conditionally independent given a set of other variables, then they do not share a direct causal link.

2. **Score-based Learning**: This method searches through the space of possible network structures by defining a scoring function to evaluate and select the best structure.
    - **Scoring Function**: This function evaluates how well a given network structure fits the data. Common scoring functions used in literature include the Akaike Information Criterion (AIC), likelihood score, and Bayesian Dirichlet equivalent uniform.
    - **Search Strategy**: In the second step, various search strategies can be utilized, such as greedy search [21], or meta-heuristic methods (e.g., genetic algorithms [65, 67], particle swarm optimization [70]) to explore the space of possible structures. and find the one that maximizes (or minimizes) the scoring function.
3. **Hybrid Learning**: Hybrid algorithms combine elements of both constraint-based and score-based approaches to achieve better performance and accuracy.
    - **Initial Constraint-based Phase**: These algorithms typically start by using constraint-based methods to reduce the search space by identifying a skeleton of the network (i.e., a graph without edge directions).
    - **Score-based Refinement**: In the next phase, a score-based method is applied to refine this skeleton and determine the directionality of the edges, ensuring a more accurate and efficient structure discovery process.

    The Max-Min Hill-Climbing (MMHC) algorithm is a remarkable example of a hybrid approach [106]. It Combines constraint-based and scoring methods. This algorithm processes the Bayesian networks in the following steps:
    - **Skeleton Identification**: The MMHC algorithm begins by identifying the skeleton of the Bayesian network, which is the undirected graph that represents the conditional dependencies between variables. This is done using a local discovery algorithm called Max-Min Parents and Children (MMPC) [16]. The MMPC algorithm seeks subsets of variables that render pairs of variables conditionally independent, effectively determining the edges of the skeleton.
    - **Edge Orientation**: After the skeleton identification, the edges are oriented to form a DAG. This is achieved through a greedy hill-climbing search guided by a Bayesian scoring function. The search is constrained to consider only those edges that were identified in the first phase. The algorithm iteratively adds, deletes, or reverses edges to maximize the score, which reflects how well the network fits the data.

    Combining the two steps helps the algorithm to learn the structure of Bayesian networks from data efficiently and the process can scale to large datasets with good accuracy.

3.4.2.1 MIIC Algorithm

The Multivariate Information based Inductive Causation (MIIC) algorithm is another hybrid structure learning method first introduced to learn from genomic data [109]. MIIC combines constraint-based and information-theoretic approaches to learn causal networks with latent variables from data to build a Bayesian Network. The MIIC algorithm integrates the strengths of constraint-based and score-based approaches to learn the network structure efficiently. The key steps involved in the algorithm are as follows:

1. **Building the Graph Skeleton**: The MIIC algorithm initiates with a fully connected graph and iteratively removes edges that are not supported by the data. It uses multivariate information to uncover indirect paths and assesses whether the mutual information between two variables can be explained by other variables in the network.

2. **Edge Filtering**: After the initial skeleton is built, edges are filtered based on their confidence levels.

3. **Edge Orientation**: The remaining edges are then oriented based on the signature of causality in observational data.

By combining these steps, the MIIC algorithm effectively learns the structure of Bayesian networks, capturing both direct and indirect dependencies between variables while accounting for potential latent variables. This makes it a robust and versatile tool for causal inference in complex datasets.

4 Method

This chapter describes the methodology applied in this thesis. Section 4.1 begins with describing the data processing on the input dataset. The second section (4.2) introduces the algorithm designed for the subsequent steps of this project. The third section 4.3 provides the two methods used for observing the trends in the large dataset, linear regression and LDA topic modeling. Following the structure of the previous chapter the last three sections (4.4,4.5,4.6) would focus on the main description of methods Vector Autoregression, Granger Causality, and Bayesian Network Analysis following by their implementation in the context of our problem. As an overview, This project involves the following steps:

1. **Data Processing:** The first step in this research is data processing. This involves preparing the dataset for analysis.
2. **Frequency Counting:** The next step is to apply a frequency counting algorithm to the 'Keywords' column. This algorithm counts the frequency of each keyword combination in the dataset. The output of this step is the frequency of occurrence for each keyword.
3. **Trends Analysis:** Linear regression and Latent Dirichlet Allocation are applied to the time series of keywords and the keywords themselves.
4. **Time Series Dataset:** Using the output of the frequency counting, a time series dataset is built.
5. **Vector Autoregression Model:**A Vector Autoregression (VAR) model is then applied to this time series dataset.
6. **Granger Causality Testing:** In parallel with the VAR model, Granger causality testing is applied.
7. **Bayesian Network:** Finally, a Bayesian Network is built for corresponding analysis.

The methods used in step 2 are tools for answering the first two questions regarding the trends of datasets, while the last three are used for the questions that correspond to influence analysis of keywords.

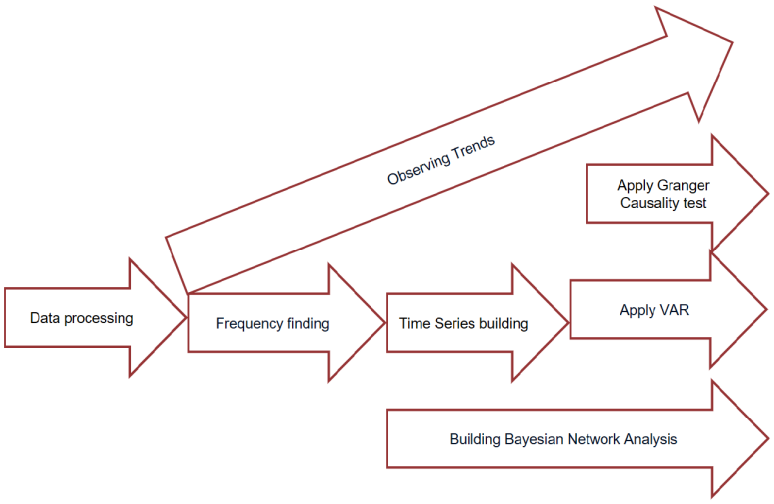


Figure 4.1: Pipeline for Research

4.1 Data preparation

This is a critical step in the thesis project. This process involves collecting, cleaning, and filling missing data from external sources. This initial step ensures that the data used in the research is accurate, consistent, and reliable.
The data was gathered from the Linköping University DiVA portal. The retrieved data includes various publications that were registered by at least one author from Linköping University.

4.1.1 Data Cleaning

The publications have 68 attributes, for the purpose of this research the following attributes are significant:

- **PID:** A unique identifier assigned to each publication, can be used as a key to distinguish between research projects.
- **Name:** This attribute contains the names of respective authors and the organization they worked with.
- **Title:** The title that the publication is published in.
- **DOI:** The DOI identifier for articles will be used for the next steps.
- **Year:** The Year of Publish relevant to the articles.
- **Keywords:** The author defined keywords for the publication.

All of the publishment understudy needs to have the corresponding keywords attribute, therefore the data-cleaning procedure involves the following steps:

1. **Name, title, and Year validation:** The three attributes should be complete and correct, this combination ensures the existence and uniqueness of the corresponding publishment. The invalid articles were removed
2. **keywords processing:** The keyword attribute contains a list of ';' separated comprising mainly of English and Swedish words in the dataset. To facilitate the algorithm execution in the next sections, the keywords were transformed to lowercase and both leading and trailing white spaces were removed.

3. **removing invalid items:** Some publishments had unknown attributes and due to DOI invalidity or unavailable DOI, recovering the accurate information seems impossible, therefore such items were flagged as invalid and removed.

After these steps, the data is passed to augmentation steps. A crucial point about the dataset is that three attributes are associated with time available: 'Year', 'CreatedDate', 'PublicationDate', and 'LastUpdated'. It was verified that the 'Year' value contains the most accurate date information and will be used in subsequent steps.

4.1.2 Data Augmentation

After data cleaning, all the items with valid keywords are assumed as valid input data, additionally, for publishments with missing or invalid keywords, a query process using a web API was performed. The Elsevier API [27] provides a rich source of information that can be used to fill in the gaps in our data. It allows us to access a wide range of scientific articles and extract relevant information such as keywords, authors, titles, and more. This process involves the following steps:

1. **API Query:** For each publication with missing or invalid keywords, we construct a query using its DOI identifier. This query is then sent to the Elsevier API.
2. **Data Extraction:** The API returns a set of results for each query. We parse these results and extract the relevant information. In this case, we are primarily interested in the author-defined keywords associated with each publication.
3. **Data Integration:** The extracted keywords are integrated into our existing dataset. This involves matching the keywords with the correct publications and updating the 'Keywords' attribute accordingly.

The updated items are validated to ensure the accuracy of the keywords.
After processing the dataset and in the next implementation of chapter 5 three subsets of the original dataset, 'Department of Computer and Information Science', 'Department of Mathematics', and 'Media and Information Technology unit' are considered.

4.2 Frequency counter algorithm

In the context of this thesis, understanding the frequency of keyword occurrences is crucial to building the time series needed for the next steps the trends and patterns in the data. The frequency-finding algorithm is designed to count the co-occurrences of keyword combinations in the dataset. The algorithm is as follows:

Algorithm 1 Keyword Co-occurrence Algorithm

```
procedure KeywordCooccurrence(filename, keyword_column, separator, A_num)
    Load the dataset from the CSV file specified by 'filename'
    for each row in the keywords of the dataset do
        Split the keywords in the row by the 'separator'
        if the row is a list then
            Strip whitespaces from each keyword and convert them to lowercase
        end if
    end for
    Initialize a Counter object to count the co-occurrences of keyword combinations
    for each list of keywords in the keywords of the dataset do
        if the list of keywords is not null then
            Find all combinations of 'A_num' keywords in the list
            Sort each combination and update their count in 'co_occurrences'
        end if
    end for
    Convert 'co_occurrences' to a list of tuples and sort it in descending order of count
    return the sorted list of tuples
end procedure
```

The algorithm takes four parameters: 'filename', $keyword_column$, 'separator', and A_num. The 'filename' specifies the CSV file containing the dataset. The $keyword_column$ indicates the column in the dataset that contains the keywords. The 'separator' is used to split the keywords in each row, and A_num specifies the number of keywords in each combination.
'Counter' is a Python [95] class that allows counting the occurrences of elements in a list. It is used here to count the occurrences of each combination of keywords.
The update method of a Counter object is used to increment the count of elements in the Counter. It is used here to increment the count of each combination of keywords.
The 'sorted' function is used to sort the list of tuples in descending order of count. Each tuple in the list contains a combination of keywords and its count.
This algorithm provides a systematic approach to count the combination objects and therefore is used in all implemented methods.

4.3 Observing Trends

4.3.1 Linear Regression for Trend Analysis

Linear regression is a statistical method that can be used to model the relationship between a dependent variable and one or more independent variables[79]. In the context of scientometrics, linear regression can be useful for identifying whether the usage of certain keywords is increasing, decreasing, or remaining stable over time.

In a simple linear regression model, the relationship between the dependent variable keyword frequency Y and the independent variable of Year X is modeled as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable (keyword frequency).
- X is the independent variable (year).
- β_0 is the y-intercept of the regression line.
- β_1 is the slope of the regression line, representing the change in keyword frequency for every one-unit change in Year.

- ϵ is the error term, representing the variation in keyword frequency that time cannot explain through linear relationships.

To estimate the parameters β_0 and β_1, we use the least squares method, which minimizes the sum of the squared differences between the observed values and the values predicted by the model.
After observing the trends corresponding to the top keywords in the large dataset, and to handle the nonlinear behavior of the values across years, the following ten distribution was tested against the top keywords:

- **Exponential**:
$$f(x;\lambda) = \lambda e^{-\lambda x}$$
Parameters:
  - λ is the rate parameter.
- **Log-Normal**:
$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$
Parameters:
  - μ is the mean of the variable's natural logarithm.
  - σ is the standard deviation of the variable's natural logarithm.
- **Power Law**:
$$f(x;k) = kx^{k-1}$$
Parameters:
  - k is the power exponent.
- **Gaussian**:
$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
Parameters:
  - μ is the mean.
  - σ is the standard deviation.
- **Rayleigh**:
$$f(x;\sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$$
Parameters:
  - σ is the scale parameter.
- **Weibull**:
$$f(x;k,\lambda) = \frac{k}{\lambda^k} (x)^{k-1} e^{-(x/\lambda)^k}$$
Parameters:
  - k is the shape parameter.
  - λ is the scale parameter.
- **Gamma**:
$$f(x;k,\theta) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}$$
Parameters:

  - k is the shape parameter.
  - θ is the scale parameter.

- **Beta**:
$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$$
Parameters:
  - α is the first shape parameter.
  - β is the second shape parameter.
  - $\Gamma(\cdot)$ is the Gamma function.

- **Chi-Square**:
$$f(x;k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}$$
Parameters:
  - k is the degrees of freedom.

- **Student's t**:
$$f(x;\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$
Parameters:
  - ν is the degrees of freedom.

To find the best values for these univariate distributions, maximum likelihood estimation for the parameters of each distribution was employed, and the resulting distribution was tested using the Kolmogorov-Smirnov test (two-sample K-S test)([91]) to compare the statistical significance of these observations. The Kolmogorov-Smirnov test statistic for a given cumulative distribution function $F(x)$ and the distribution of keywords, is defined as:

$$D_n = \sup_x |F_n(x) - F(x)|$$

- $\sup_x$ is the supremum of the set of distances.
- $|F_n(x) - F(x)|$ is the absolute difference between the keyword data and the theoretical cumulative distribution function.

The null hypothesis states that the keyword distribution follows the theoretical distribution, and is rejected at level α if:

$$D_n > c(\alpha)\sqrt{\frac{n}{N_n}}$$

- $c(\alpha)$ is the inverse of the Kolmogorov distribution at $1-\alpha$.
- n is the number of keywords.
- N_n is the number of keywords the random variable can take on.

4.3.2 Latent Dirichlet Allocation

To identify the hidden topic structure based on the keywords to capture the thematic content of the documents, there is a need to infer the hidden variables given the observed data3.1.3. This is done using various approximation approaches such as Gibbs Sampling[40] and Collapsed variational inference[104]. The method employed in this thesis is a variant of variational inference called Online Variational Bayes (VB)([47]), this method can handle massive document collections effectively incrementally for large-scale datasets.
Variational Bayes approximates the true posterior distribution $p(\theta, \mathbf{z}, \beta|\mathbf{w}, \alpha, \eta)$ with a simpler distribution $q(\theta, \mathbf{z}, \beta|\lambda, \phi, \gamma)$ as close as possible to the true posterior by minimizing the Kullback-Leibler (KL) divergence between them. This is equivalent to maximizing the Evidence Lower Bound (ELBO). In the case of LDA, the variational distribution is factorized as:

$$q(\theta, z, \pi|\lambda) = q(\pi|\delta) \prod_{i=1}^{N} q(z_i|\phi_i) \prod_{i=1}^{M} q(\theta_i|\gamma_i)$$

Here, λ, ϕ, and δ are variational parameters. Furthermore, the ELBO can be expressed as follows:

$$\begin{aligned} L = & \sum_d \sum_w n_{dw} \sum_k \phi_{dwk} (Eq[\log \theta_{dk}] + Eq[\log \beta_{kw}] - \log \phi_{dwk}) \\ & - \log \Gamma(\sum_k \gamma_{dk}) + \sum_k (\alpha - \gamma_{dk}) Eq[\log \theta_{dk}] + \log \Gamma(\gamma_{dk}) \\ & + \frac{\sum_k - \log \Gamma(\sum_w \lambda_{kw}) + \sum_w (\eta - \lambda_{kw}) Eq[\log \beta_{kw}] + \log \Gamma(\lambda_{kw})}{D} \\ & + \log \Gamma(K\alpha) - K \log \Gamma(\alpha) \\ & + \frac{\log \Gamma(W\eta) - W \log \Gamma(\eta)}{D} \end{aligned}$$

Where:

- L is the ELBO, which needs to be maximized.
- d and w are indices for documents and words, respectively.
- n_{dw} is the number of times word w appears in document d.
- ϕ_{dwk} is the variational parameter for the topic assignment of word w in document d to topic k.
- θ_{dk} is the topic distribution for document d.
- β_{kw} is the word distribution for topic k.
- γ_{dk} is the variational parameter for the topic distribution of document d.
- λ_{kw} is the variational parameter for the word distribution of topic k.
- α and η are hyperparameters of the Dirichlet priors on the topic and word distributions, respectively.
- K is the number of topics, and W is the size of the vocabulary.
- D is the number of documents.
- Γ is the gamma function, which generalizes the factorial function to real numbers.

The goal is to adjust the variational parameters (ϕ, γ) to maximize this quantity, which results in the variational distributions closely approximating the true posterior distributions of the hidden variables in the model. This is done by updating the following expressions:

$$\phi_{dwk} \propto \exp\{\mathrm{Eq}[\log \theta_{dk}] + \mathrm{Eq}[\log \beta_{kw}]\}$$

$$\gamma_{dk} = \alpha + \sum_{w} n_{dw} \phi_{dwk}$$

In this Thesis, the collection of keywords is treated as documents and LDA topic modeling was applied according to the implementation of Online variational Bayes with gensim [93] library. Gensim handles large datasets efficiently by performing the algorithm on documents in mini-batches, the overall structure of this process is as follows:

1. Initialize λ randomly.
2. For each mini-batch of documents.
    - Perform an E-step to update ϕ and γ for the mini-batch.
    - Compute a sufficient statistics update for $\hat{\lambda}$
    - Update λ using a weighted average of the old λ and the sufficient statistics.

Following the results, three metrics are used to evaluate the LDA model:

1. **Perplexity**: It measures how well a probabilistic model predicts a sample. In this application, it evaluates the overall quality of topics containing keywords. A lower perplexity score indicates better generalization performance. The perplexity of a corpus D as a list of keywords given an LDA model with parameters α and β is defined as:

    $$\text{Perplexity}(D) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d|\alpha, \beta)}{\sum_{d=1}^{M} N_d} \right\}$$

    Where:

    - M is the number of documents.
    - N_d is the number of keywords in document d.
    - $p(w_d|\alpha, \beta)$ is the likelihood of the words in document d given the LDA model parameters.

    as an alternative form for this equation, the lower bound for perplexity is computed:

    $$\text{perplexity } (\boldsymbol{n}, \boldsymbol{\lambda}, \alpha) \leqslant \exp\left\{ -\left(\sum_{d} \mathbb{E}_q \left[\log p\left(n_d, \theta_d, z_d \mid \alpha, \boldsymbol{\beta}\right)\right] - \mathbb{E}_q\left[\log q\left(\theta_d, z_d\right)\right] \right) \left(\sum_{d,w} n_{dw} \right) \right\}$$

    similar to the original form the smaller values(larger negatives are more desirable).

2. **Topic Difference (Jaccard Distance)**: This metric shows the dissimilarity between two topics. For every pair of topics (A, B), the Jaccard distance is defined as:

    $$\text{Jaccard Distance}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

    This is applied to the sets of top words in each topic to measure how different the topics are. As a high dimensional metric, the difference is reported in the form of heatmaps containing pairwise comparisons for every combination of topics

3. **Hellinger Distance**: The last metric, is a measure to quantify the similarity between two probability distributions. For every pair of topics $P = (p_1, p_2, \ldots, p_k)$ and $Q = (q_1, q_2, \ldots, q_k)$, the Hellinger distance is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

Similar to the previous metric, Hellinger Distance is computed for all topics after setting the hyperparameter for the number of topics, however, unlike the Jaccard distance where the topics are compared with respect to groups of keywords, this metric considers the distribution of topics in documents. Hence, for every instance of the model, boxplots were employed to visualize the distributions per iteration effectively.

4.4 Vector Autoregression

The first methodology used in this thesis is Vector Autoregression (VAR)in order to capture the linear interdependencies among multiple time series data. In the context of scientometrics, each element of the vector Y_t represents the presence of a specific keyword within the scientific literature at time t. By applying the VAR model, we can analyze how previous occurrences of keywords influence the current state of other keywords and uncover the dynamic relationship among them over time. considering a simplified VAR(1) model with two keywords, the model can be expressed as:

$$\textbf{Keyword1}_t = \alpha_{11}\textbf{Keyword1}_{t-1} + \alpha_{12}\textbf{Keyword2}_{t-1} + \varepsilon_{1t}$$

$$\textbf{Keyword2}_t = \alpha_{21}\textbf{Keyword1}_{t-1} + \alpha_{22}\textbf{Keyword2}_{t-1} + \varepsilon_{2t}$$

where α_{ij} represents the influence of keyword j at time $t-1$ on keyword i at time t, and ε_{it} denotes the error term for keyword i at time t. T In the context of this thesis, the VAR model is applied To implement this model using the provided data the Akaike Information Criterion (AIC) is used to determine the optimal lag length for the keyword combination series.

4.5 Causality and Granger Causality test

Causality primarily involves the cause-and-effect relationship in a phenomenon [88]. In scientific research, establishing causality requires a stronger relationship than simple correlation so that the changes in **Causing** variable lead to changes in the **effect**. In the context of keyword analysis in this thesis, causality refers to understanding the changes in the occurrence of one keyword influencing changes in another keyword. Mathematically, if K_1 and K_2 represent two keywords, the average causal effect of K_1 on K_2 can be defined as:

$$\text{ACE}(K_1 \to K_2) = \mathbb{E}[K_2 \mid \text{do}(K_1 = 1)] - \mathbb{E}[K_2 \mid \text{do}(K_1 = 0)]$$

For implementation of the algorithms, the following definitions are given in the context of keyword analysis

4.5.1 Directed Acyclic Graph (DAG)

A Directed Acyclic Graph (DAG) is a finite graph with directed edges and no cycles. They represent causal relationships among keywords. Each node in the DAG represents a keyword, and a directed edge from keyword A to keyword B indicates that the presence of A influences the presence of B. Mathematically, a DAG $G = (V, E)$ consists of:

- A set of vertices V (variables).
- A set of directed edges E where $E \subseteq V \times V$.

A path is a sequence of edges that connect a sequence of vertices. According to the definition above, no path starts and ends at the same vertex in a DAG.

4.5.2 Confounding Variable in Keyword Analysis

A confounding variable is a keyword that affects both the cause and the effect keywords, potentially leading to incorrect conclusions about their relationship.
Example: In studying the relationship between "big data" and "cloud computing":

- A confounding keyword might be "data mining". Both "big data" and "cloud computing" might be discussed in connection with "data mining".
- The DAG might show:
  - Data Mining → Big Data
  - Data Mining → Cloud Computing
- Failing to account for "data mining" will falsely suggest a direct causal relationship between "big data" and "cloud computing".

4.5.2.1 Back-door Path

A back-door path involves an indirect path that can create confounding.
Example:

- "Artificial Intelligence" (*AI*) and "Robotics" with a confounding variable "Computer Vision":
  - AI ← Computer Vision → Robotics

4.5.2.2 Front-door Path

A front-door path involves a mediator keyword through which the causal influence flows.
Example:

- "Neural Network" (NN) influences "Deep Learning" (DL) through "Machine Learning" (ML):
  - NN → ML → DL

4.5.2.3 Back-door Criterion

A set of keywords Z satisfies the back-door criterion for X and Y if:

1. No keyword in Z is a descendant of X, so that none of the keywords in the set Z are influenced by X directly or indirectly.
2. The set Z must block all paths from X to Y that contain an arrow pointing into cause X. By blocking these paths, Z prevents any potential confounding variables from influencing the observed relationship between X and Y.

4.5.2.4 Front-door Criterion

A set of keywords Z satisfies the front-door criterion if:

1. Z should intervene in any pathways that lead from X to Y through other keywords, ensuring that the observed relationship between X and Y is not confounded by other factors.
2. After removing X, there should be no paths from Z to Y that contain an arrow pointing away from Z.

4.5.3 Do-Calculus

Do-Calculus is a set of rules developed by Judea Pearl[108] for reasoning about causal effects through interventions in a probabilistic framework. In mathematical terms, do-calculus involves the manipulation of expressions in the form of

$$P(Y \mid do(X))$$

, represents the probability distribution of Y when X is set by an intervention. The Do-Calculus includes three key rules:

1. **Insertion/Deletion of Observations**

   Rule:

   $$P(Y \mid do(X), Z, W) = P(Y \mid do(X), W) \quad \text{if } Y \perp Z \mid X, W \text{ in } G_{\overline{X}}$$

   Explanation:

   - First rule excludes an observed variable Z from the conditional distribution if Y is independent of Z given X and W after intervening on X.
   - The graph $G_{\overline{X}}$ represents the scenario where all incoming edges to X are removed, signifying an intervention on X.

2. **Action/Observation Exchange**

   Rule:

   $$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), Z, W) \quad \text{if } Y \perp Z \mid X, W \text{ in } G_{\overline{X},\underline{Z}}$$

   Explanation:

   - Second rule replaces an intervention on Z with an observation of Z if Y is independent of Z given X and W in the graph.
   - The graph $G_{\overline{X},\underline{Z}}$ represents the scenario where all incoming edges to X and all outgoing edges from Z are removed, indicating a state of intervention on X.

3. **Insertion/Deletion of Actions**

   Rule:

   $$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), W) \quad \text{if } Y \perp Z \mid X, W \text{ in } G_{\overline{X},\overline{Z(W)}}$$

   Explanation:

   - Third rule allows to exclude an intervention on Z from the conditional distribution if Y is independent of Z given X and W in the graph.
   - Given Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$, the graph $G_{\overline{X},\overline{Z(W)}}$ represents the scenario where all incoming edges to X and $Z(W)$ are removed, indicating interventions on X and Z.

4.5.4 Granger Causality in Keyword Analysis

Granger causality is a statistical hypothesis test that determines if one time series can predict another[97]. In the context of this thesis, Granger causality tests whether the frequency of one keyword (e.g., "linear algebra") can predict the frequency of another keyword (e.g., "matrix theory") over time.

Traditional causality methods, such as those involving Directed Acyclic Graphs (DAGs) and do-calculus, focus on identifying and understanding causal relationships through a network of variables and interventions. These methods are powerful for uncovering direct and indirect causal effects and for handling confounding variables. However, they often do not account for temporal dependencies explicitly.

In contrast, Granger causality is specifically designed to handle time series data, making it particularly useful for understanding how the historical values of one keyword can influence the future values of another. Another key point is the incorporation of time lags, which allows for the detection of predictive relationships that unfold over time.
Consider two time series of keyword frequencies X_t (e.g., "linear algebra") and Y_t (e.g., "matrix theory"). To test whether "linear algebra" Granger-causes "matrix theory":

1. **Model 1 (without X):**
$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \epsilon_t$$
2. **Model 2 (with X):**
$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{q} \gamma_j X_{t-j} + \epsilon_t$$

To determine if "linear algebra" Granger-causes "matrix theory," we perform an F-test to check if the coefficients γ_j are significantly different from zero in Model 2 for a selected number of time lags. If they are, it indicates that past values of "linear algebra" help predict "matrix theory" in each chosen time lag, implying Granger causality.

4.6 Bayesian Network Analysis

A Bayesian network is a graphical model that consists of a set of nodes (representing variables) and directed edges (representing the relationships between the variables). The structure of the network encodes the conditional dependencies between the variables, and the parameters of the network (i.e., the conditional probability distributions) are learned from data or expert knowledge.

In this project, two approaches were applied to constructing the network from data.
In the subsequent chapters 'Expert-Driven Learning' and the model based on 'Structure learning' will be known as 'designed network' and 'learned network' respectively.

4.6.1 Expert-Driven Learning

In the first approach, the network structure is specified based on expert knowledge. It is assumed that the presence of keywords provides a strong prior understanding of the overall structure. The set of all publications denoted as $P = \{p_1, p_2, ..., p_n\}$ and the set of all unique keywords extracted from these publications as $K = \{k_1, k_2, ..., k_m\}$.

1. **Node Creation**: Using the extracted keywords from Frequency counter(4.2), Each keyword k_i in K is used to create a node in the Bayesian network. The state of each node X_i can be either 0 (keyword not present) or 1 (keyword present). therefore a binary dataset was created. Each node represents a keyword and its state represents whether the keyword is present (1) or absent (0) in a particular publication.

2. **Conditional Probability Distributions (CPDs)**: For each node X_i, we define a CPD $P(X_i|Pa(X_i))$, where $Pa(X_i)$ denotes the parents of node X_i in the network. This CPD represents the probability of the presence or absence of keyword k_i given the presence or absence of its parent keywords. These probabilities can be estimated from the binary dataset where each row represents a publication and each column represents a keyword (node), with 1s and 0s indicating the presence or absence of the keyword in the publication.

3. **Network Construction**: The Bayesian network is represented as a Directed Acyclic Graph (DAG) $G = (V, E)$, where V is the set of nodes (representing the keywords) and E is the set of directed edges (representing the conditional dependencies between the keywords). An edge from node X_i to node X_j in the network indicates that the presence or absence of keyword k_i has a direct influence on the presence or absence of keyword k_j.

   The structure of the Bayesian network, or the set of edges E, can be represented by an adjacency matrix A of size $m \times m$. If there is a directed edge from node X_i to node X_j, then $A_{ij} = 1$; otherwise, $A_{ij} = 0$.

   The parents of a node X_i, denoted as $Pa(X_i)$, are the nodes that have a directed edge to X_i. Mathematically, $Pa(X_i) = \{X_j : A_{ji} = 1\}$.

   The children of a node X_i, denoted as $Ch(X_i)$, are the nodes that X_i has a directed edge to. Mathematically, $Ch(X_i) = \{X_j : A_{ij} = 1\}$.

   The joint probability distribution over all nodes $X_1, ..., X_m$ in the network is given by the product of the CPDs of all nodes:

$$P(X_1, ..., X_m) = \prod_{i=1}^{m} P(X_i|Pa(X_i))$$

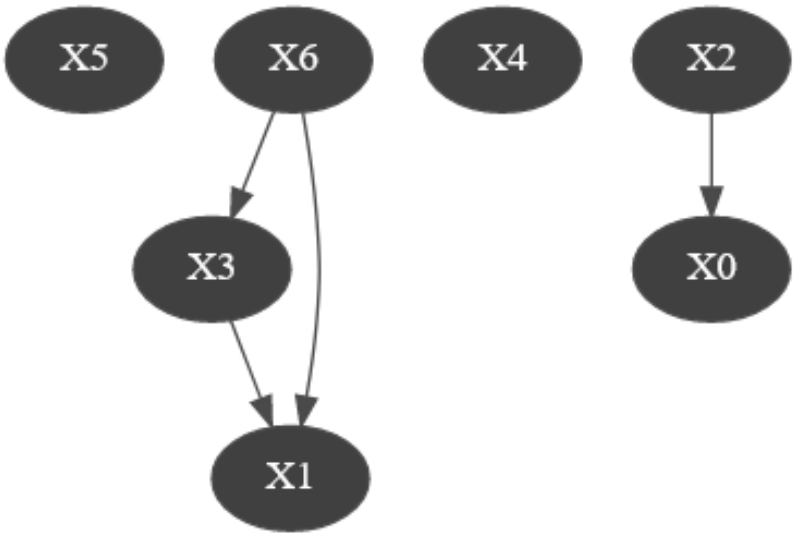


Figure 4.2: Network structure on Random Data

For the three nodes the joint distribution would be:

$$P(X_1, X_3, X_6) = P(X_6) \cdot P(X_1|X_3, X_6) \cdot P(X_3|X6)$$

The joint distribution can be used to answer various probabilistic queries about the key-

Table 4.1: Probability of occurrences for variables

	X0	
X2	0	1
0	0.9997	0.0003
1	0.9998	0.0002

X2	
0	1
0.5000	0.5000

X4	
0	1
0.4880	0.5120

words, such as the probability of a keyword appearing in a publication given the presence of other keywords. It can also be used to identify which keywords tend to appear together in publications or to predict the likelihood of a keyword appearing in a publication given the presence of other keywords.

4.6.2 Structure Learning

The goal of structure learning is to infer the underlying network structure from the data, which can be used for various applications such as predicting the probability of a particular outcome, identifying the causal relationships between variables, and making decisions under uncertainty. Hybrid structure learning algorithms have been developed to address the limitations of traditional methods. These algorithms combine different approaches, such as Bayesian model averaging and greedy searches, to provide a collection of samples from the posterior distribution of the graph given the data. This allows for Bayesian model averaging, which is particularly useful in high-dimensional domains with sparse data, where a single best structure cannot be identified from the data. The MIIC algorithm is based on the idea that the mutual information between two variables, conditioned on a set of other variables, can be used to infer the causal relationships between them. This approach is motivated by the fact that mutual information can capture the statistical dependencies between variables, which are a key aspect of causal relationships.

1. **Initialization**: Start with a complete undirected graph $G = (V, E)$, where V is the set of nodes (representing the keywords) and E is the set of edges (initially connecting all pairs of nodes). If we have m unique keywords, then $V = \{X_1, X_2, ..., X_m\}$ and $E = \{(X_i, X_j) : i, j \in \{1, 2, ..., m\}, i \neq j\}$.

2. **Building the Graph Skeleton**: The MIIC algorithm starts by iteratively removing edges that are not supported by the data. For each edge $e = (X_i, X_j) \in E$, calculate the total information contribution $I(X_i; Y_i; \{A_j\})$ of all indirect paths between the nodes of e. The iterative decomposition of mutual information is given by:

$$I(X_i; X_j) = \sum_{k=1}^{n} I(X_i; X_j; A_k \mid \{A_l\}_{l=1}^{k-1}) + I(X_i; X_j \mid \{A_k\}_{k=1}^{n})$$

   where $I(X_i; X_j; A_k \mid \{A_l\}_{l=1}^{k-1})$ is the mutual information contribution of the k-th variable conditioned on the previously collected variables. If $I(X_i; X_j \mid \{A_k\}_{k=1}^{n}) \approx 0$, then X_i and X_j are conditionally independent given A_k, and the edge between X_i and X_j is removed.

3. **Edge Confidence**: After building the initial skeleton, calculate the confidence $C(X_i, X_j)$ of each remaining edge $e = (X_i, X_j) \in E$ based on randomization of the data. The confidence of an edge can be computed as the proportion of randomized datasets for which the information contribution $I(X_i; X_j)$ is less than or equal to the observed $I(X_i; X_j)$:

$$C(X_i, X_j) = \frac{P_{X_i, X_j}}{[P^{\text{rand}}_{X_i, X_j}]}$$

where P_{X_i,X_j} is the probability to remove edge X_iX_j and $[P^{\text{rand}}_{X_i,X_j}]$ is the average probability after randomizing the dataset.

4. **Edge Orientation**: For each remaining edge $e = (X_i, X_j) \in E$, determine the direction of e based on the signature of causality in the data. This involves identifying v-structures (a specific pattern of connections) and using them to infer the direction of causality. V-structures are identified by the presence of negative conditional mutual information:

$$I(X_i; X_j; A_n \mid \{A_k\}_{k=1}^{n-1}) < 0$$

which indicates a causal relationship such as $X_i \rightarrow A_n \leftarrow X_j$. Orient the edges accordingly and propagate these orientations throughout the network.

5. The resulting graph $G = (V, E)$ is a Bayesian network that represents the learned structure of the data. The joint probability distribution over all nodes $X_1, ..., X_m$ in the network is given by the product of the conditional probability distributions (CPDs) of all nodes:

$$P(X_1, ..., X_m) = \prod_{i=1}^{m} P(X_i \mid \text{Pa}(X_i))$$

where $\text{Pa}(X_i)$ denotes the parents of node X_i in the network.

Comparing 4.2 with previous results in 4.1 shows significant differences. The primary reason is the mathematical formulation of the MIIC algorithm, which is based on the concept of mutual information, a measure of the statistical dependence between two random variables. The algorithm uses mutual information to identify the most likely edges in the graph, orient the edges based on the conditional independence relationships between variables, and propagate the orientations of v-structures to as many remaining undirected edges as possible.

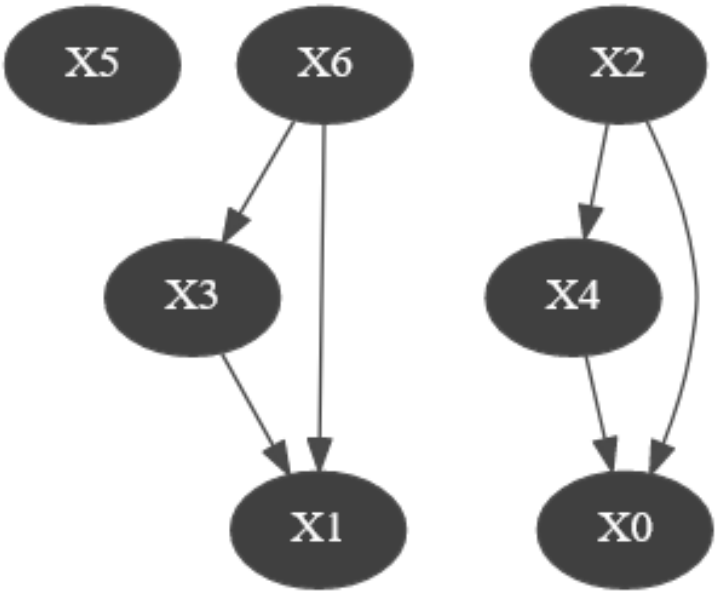


Figure 4.3: Network structure on Random Data with MIIC

Table 4.2: Probability of occurrences for variables with MIIC

		X0	
X2	**X4**	**0**	**1**
0	**0**	0.9997	0.0003
	1	0.9998	0.0002
1	**0**	0.0002	0.9998
	1	0.0003	0.9997

5 Results

This chapter describes the results applied in this thesis for answering the thesis questions. To better describe the results this chapter is divided into 4 sections corresponding to the large dataset comparison of all publications, two subsets of the Media and Information Technology division, Department of Computer and Information Science. Section 5.1 answers the first two questions. The second section 5.2 focuses on the results of the keywords relationship of the two abovementioned entities(5.2.1,5.2.2 addresses the questions concerning keywords influences.

5.1 Observed Trends

After processing the data, there are 55,980 records among the publishments that have at least one keyword. According to the dataset for any publication:

- The total number of unique keywords is 26,781.
- The minimum number of keywords an article has is 1
- The maximum number of keywords is 170(which is an outlier).
- The mean and standard deviation are 5.21 and 3.07 respectively.
- The quantiles for [25%,50%, 75%] are [4, 5, 6]

Top keywords are extracted using 4.2 and for this analysis, the top 40 keywords in each dataset and the large dataset are considered. The linear regression model provides a simple and interpretable summary of the relationship between the independent variable (time) and the dependent variable (keyword frequency), the details of this model for the top 10 words are available in table 5.1. Additionally, the equation defined for this regression model is as follows:

$$y = mx + b$$

where:

- y is the keyword frequency and x is the one unit of time (Year).

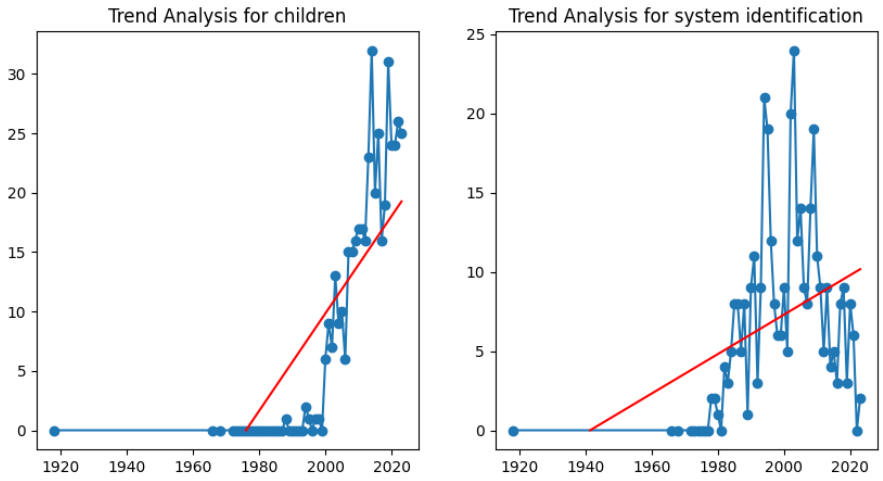


Figure 5.1: Regression Lines for children and system identification keywords

- **Slope (m):** The slope of the regression line indicates the average change in the frequency of the keyword per unit of time. A positive slope indicates an increasing trend, while a negative slope indicates a decreasing trend. The magnitude of the slope represents the rate of change.
- **R-squared (R^2):** The coefficient of determination (R^2) measures the goodness of fit of the regression model. It indicates the proportion of the variance in the dependent variable that is explained by the independent variable (time).
- **P-value:** The p-value associated with the slope tests the null hypothesis that the slope is equal to zero (i.e., there is no trend).
- **Standard Error:** The standard error of the slope estimates the variability of the slope parameter. A smaller standard error indicates greater precision in the estimation of the slope.

Among the top 40 keywords, no negative value for the slope was observed. The most interesting keywords according to this model are the ranked 28th 'type 1 diabetes' with a frequency of 201, R_2 of 0.542, and error of 1.033e-10 and also the ranked 4th 'heart failure' with a frequency of 484, R_2 of 0.497, and standard error of 0.072. For some healthcare-related keywords including these, it is observed that after the year 2015, the time series has a minimum level of occurrence, which indicates that there is a constant flow for the presence of such keywords within the dataset. Figure 5.1 shows an example of these drawn plots.

According to the plots, the keywords generally do not fall on a linear line; therefore, to interpret the results, 10 distributions were tested. Considering the previous example of ranked 4th 5.2 and 5th 5.3 as above, based on the Kolmogorov-Smirnov test at 95% confidence level, the two identified distributions and lognormal and power law are rarely the distributions of data(5.2,5.3) and are rejected. This pattern holds for many other keywords within the datasets, while the likely candidates for ma and beta distributions. Another interesting pattern that was observed is that almost always, the candidate distributions with the least p-values have a higher difference between the mean values of the data generated from that distribution and the real data from the large dataset, the reverse is not true although in many cases those with higher p-values has medium differences between mean values(rank 3 to 6). Similar patterns exist for the differences of standard deviations between the real data and

Table 5.1: Top 10 frequent keywords in the dataset.

Keywords	Count	Slope	R-squared	P-Value	Standard Error
Sweden	844	0.741	0.566	2.299e-11	0.088
gender	675	0.542	0.313	7.292e-06	0.109
sverige	609	0.508	0.490	1.904e-09	0.071
heart failure	484	0.497	0.467	6.498e-09	0.072
children	430	0.386	0.562	3.028e-11	0.046
system identification	367	0.110	0.125	0.008	0.040
depression	330	0.332	0.473	4.763e-09	0.048
optimization	328	0.234	0.533	1.767e-10	0.030
education	323	0.253	0.402	1.577e-07	0.042
quality of life	320	0.303	0.533	2.514e-10	0.038

generated data, although the rank for these differences does not always coincide. For more complete data please refer to 8.1. If we sort the result based on the highest p-value, there are 23 keywords out of the original top 40 in the interval of (0.41,0.99) and approach among these the prevalent distribution is student's t distribution closely followed by Rayleigh, for all the top keywords, $\frac{10}{40}$ follow t distribution, namely [visualization, stress, quality of life, system identification, education, simulation, communication, parameter estimation, innovation, genus]. If we expand the results to the top 500 keywords, the top candidate distribution will remain student's t with 169 value and the Gaussian distribution will be the second rank with 122 value. Do note that the values only indicate the top-selected distribution for each keyword according to p-values, in fact for some keywords the distance between the top two suitable distributions is not large, for an instance for the keyword 'COVID-19' in the dataset with a p-value of 0.9921 for the Rayleigh distribution and a p-value of 0.9772 for t distribution we can't prioritize one over the other.

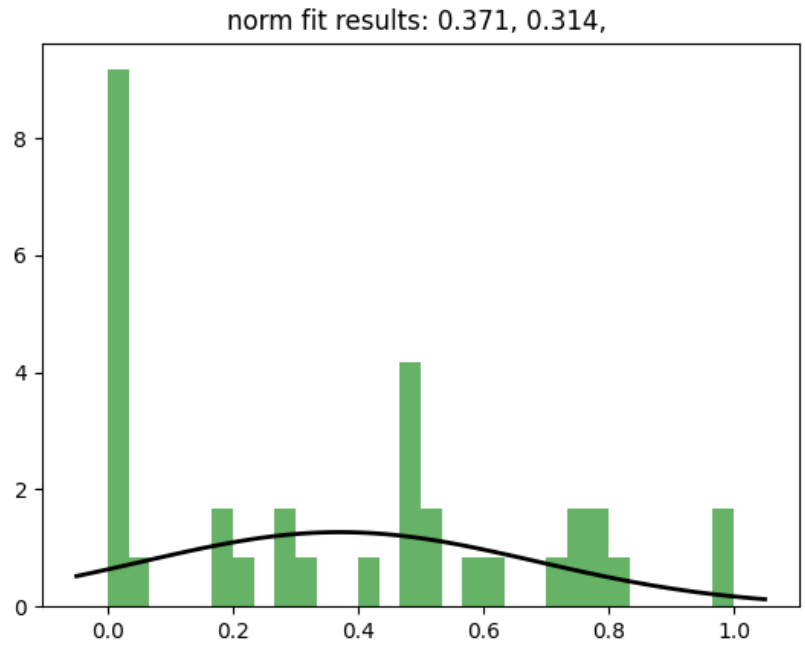


Figure 5.2: Fitting the Gaussian distribution to the Children keyword data alongside parameters

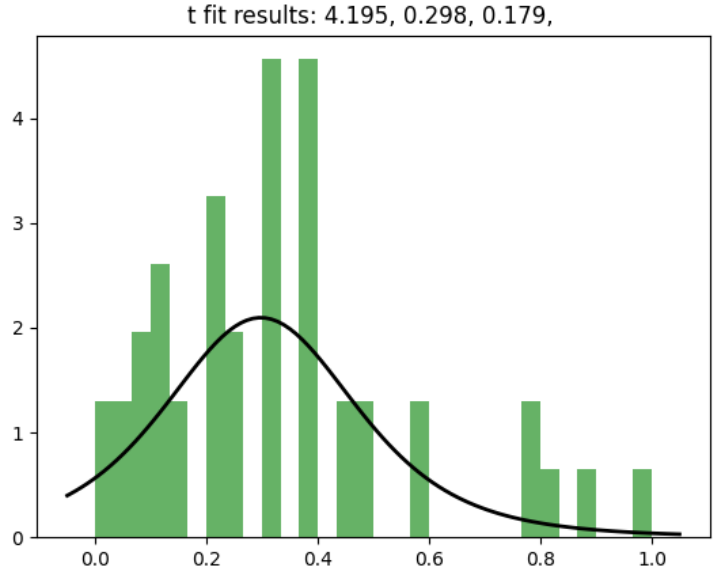


Figure 5.3: Fitting the T distribution to the Children keyword data alongside parameters

Table 5.2: Fitted distributions for Children's keyword at 95% confidence level(higher ranks indicate larger difference)

Distribution	D-stat	p-value	Mean Rank	Std Dev Rank
powerlaw	0.780874	9.52E-24	9	9
gamma	0.572495	1.07E-11	8	8
lognorm	0.49117	1.55E-08	10	10
beta	0.298192	0.002469	6	7
chi2	0.232075	0.034574	5	2
expon	0.224703	0.044477	1.5	3
rayleigh	0.199879	0.097695	4	6
weibull_min	0.194444	0.114617	7	1
t	0.169817	0.223412	3	4
norm	0.169815	0.223427	1.5	5

Table 5.3: Fitted distributions for system identification's keyword(higher ranks indicate larger difference)

Distribution	D-stat	p-value	Mean Rank	Std Dev Rank
powerlaw	0.862094	5.96E-40	9	9
lognorm	0.536089	6.92E-13	10	10
beta	0.224545	0.016122	8	2
norm	0.188984	0.065209	1.5	7
expon	0.183041	0.080422	1.5	1
rayleigh	0.162385	0.158086	6	8
chi2	0.132066	0.366387	3	5
gamma	0.132062	0.366428	4	4
weibull_min	0.12392	0.444332	5	6
t	0.117804	0.508309	7	3

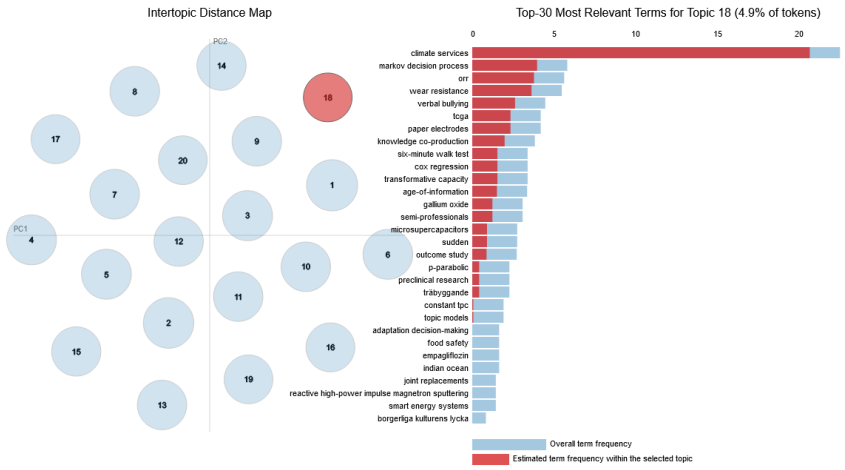


Figure 5.4: LDA with 40 topics - Relevant papers in topics as well as intertopic Distance MAP

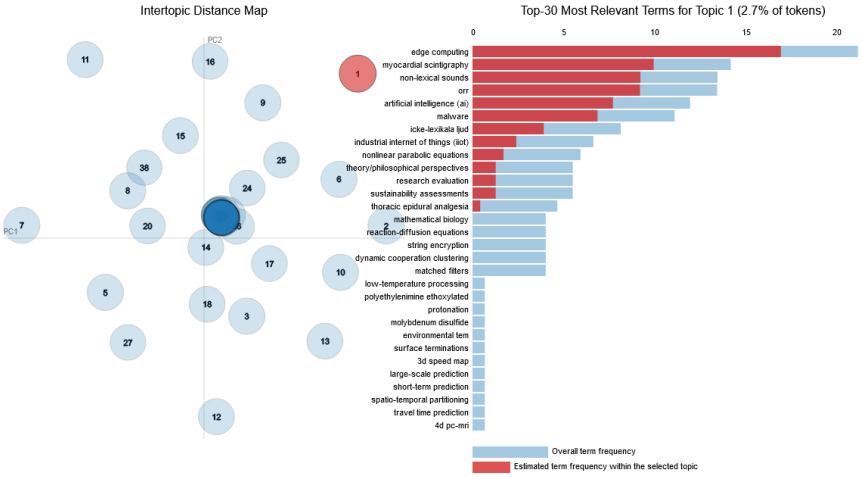


Figure 5.5: LDA with 40 topics - Relevant papers in topics as well as intertopic Distance MAP

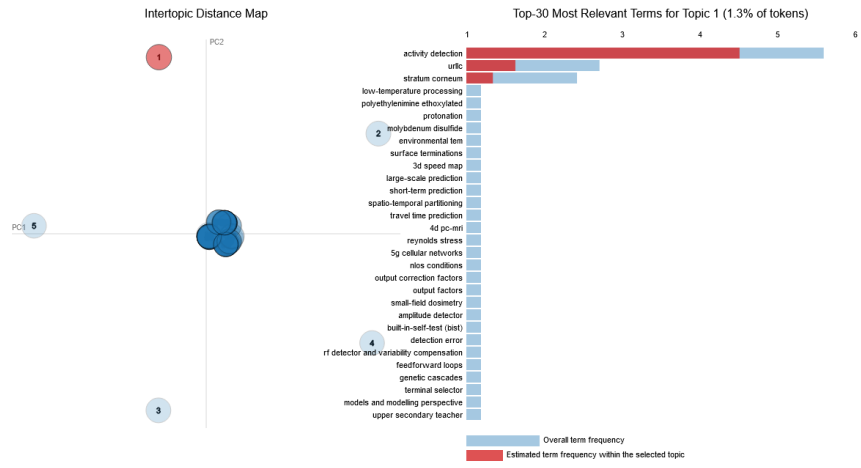


Figure 5.6: LDA with 80 topics - Relevant papers in topics as well as intertopic Distance MAP

On the other hand, we try to discover the hidden theme by utilizing the Latent Dirichlet Allocation algorithm. By analyzing the large dataset for the past five years(2019-2023) and only the keywords that emerged during this period we constructed. The LDA algorithm in this section is trained for 10 epochs with a maximum of 5000 iterations for inferring topic distributions. In this operation, the LDA was performed on a limited list of topic numbers due to the size and ultimately the expensiveness of this algorithm namely [5,10,20,30,40,50,60,70,80,90,100].

1. According to 5.7 The Per-Word Likelihood Bound shows that the gradual increase in the number of topics would increase the performance of the model in fitting to the data. For this list of topic numbers, the likelihood shows a near-linear trend, and the equation for this line is as follows:

$$y = -2.0086 \times x + (-0.2383)$$

2. 5.8 shows the Jaccard difference between the discovered topic. In the beginning, there are no similarities between the topics constructed and the keywords contained in every topic are unique, increasing the number of clusters has a direct effect on increasing the overlap between the clusters. Hence, selecting 40 as the number of topics would generate two identical clusters.

3. 5.9 shows the distribution of Hellinger Distance between each pair of clusters. Among the observed list of values, the min value is 0, the max value is approximately 0.6. By looking at the box plots, two interesting details can be noted:

   - Firstly, mean and median values are very close for each iteration, suggesting that the data is symmetrically distributed. In other words, the data is evenly distributed on both sides of the median. This is a characteristic of a Gaussian distribution.
   - Although the model's convergence to similarity follows an almost linear pattern, the pattern for outliers is different. In other words, the dissimilarity between some topics within each iteration does not follow a predictable pattern.

   In these iterations, the standard deviation fluctuates between 0.01 to 0.4. A higher standard deviation suggests a greater variability in the dissimilarity between topics in the corresponding iterations. The data for these iterations can be observed in table 5.4.

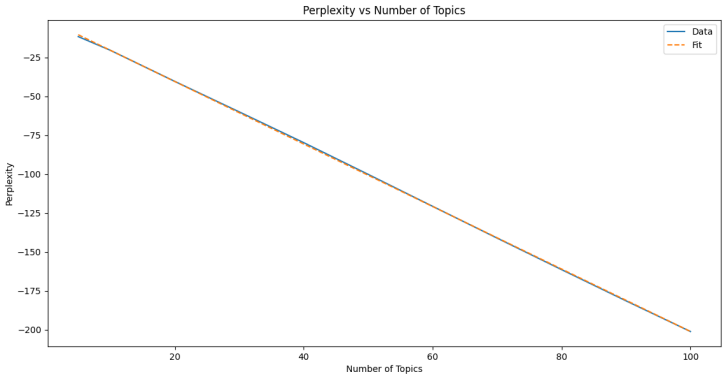


Figure 5.7: Per-Word Log Likelihood Bound (Negative Values) for the number of topics between 5 to 100(list of values)

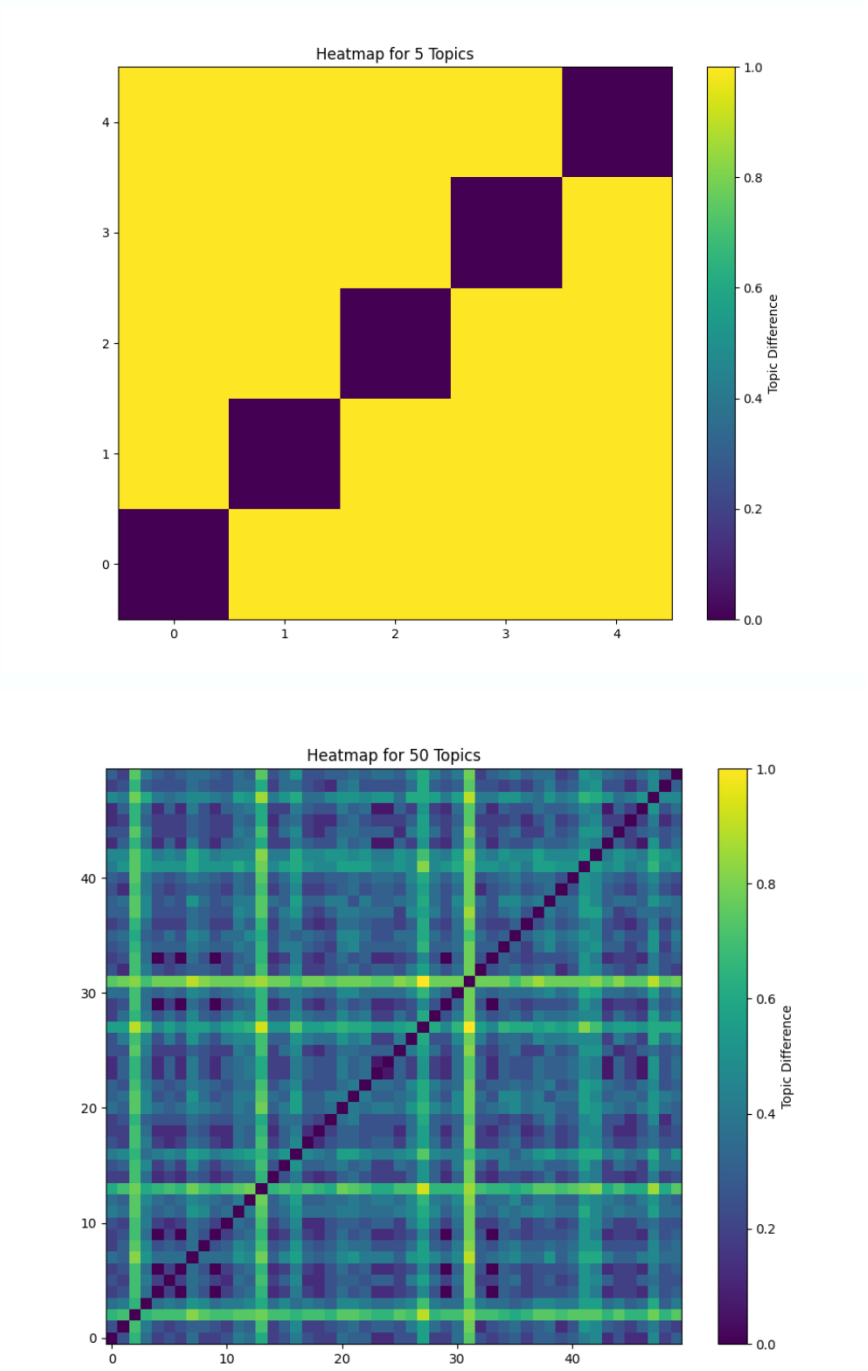


Figure 5.8: topic differences with Jaccard distance measure

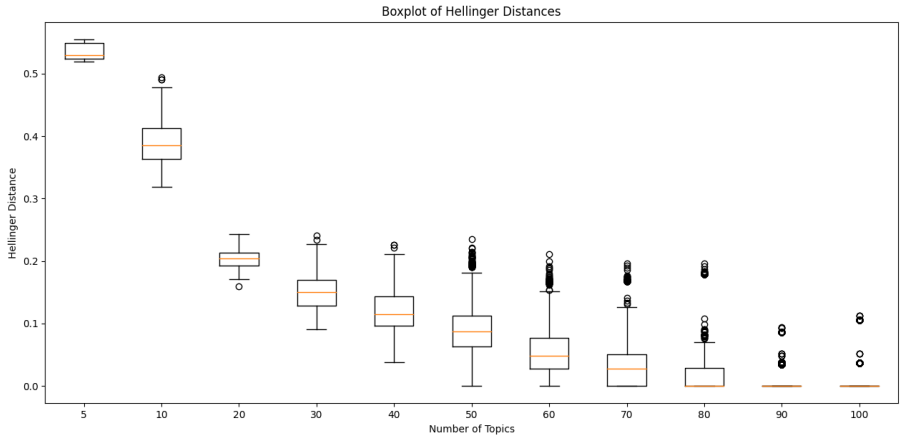


Figure 5.9: Hellinger Distance for the number of topics between 5 to 100(list of values)

Table 5.4: List of iterations for large dataset discovered topics according to Hellinger distance.

num_topics	min	max	mean	median	std_dev	1st_quantile	2nd_quantile	3rd_quantile
5	0.519135	0.554955	0.535707	0.529709	0.013341	0.524413	0.529709	0.549662
10	0.31898	0.494389	0.39556	0.385164	0.046213	0.363691	0.385164	0.412445
20	0.159053	0.243518	0.203287	0.204507	0.015977	0.192468	0.204507	0.213572
30	0.090167	0.24121	0.151456	0.150059	0.029072	0.128486	0.150059	0.16945
40	0.037528	0.226106	0.120888	0.114829	0.034145	0.096464	0.114829	0.143087
50	0	0.235539	0.090259	0.08679	0.03886	0.062851	0.08679	0.111972
60	0	0.210869	0.054951	0.047702	0.043322	0.027365	0.047702	0.077085
70	0	0.196171	0.031721	0.027054	0.039652	0	0.027054	0.050802
80	0	0.195773	0.015352	0	0.03391	0	0	0.029178
90	0	0.094592	0.005004	0	0.015896	0	0	0
100	0	0.112027	0.004274	0	0.016913	0	0	0

The results of topic modeling are as follows (considering the top 30 words in each topic):

1.
   - model with 20 number of topics and the negative Log Likelihood Bound of -40.3305. One topic in this model is as follows:
   - $"braingraph" + "digestate" + "copula" + "designspaceexploration" + "crudeoil" + "schoolchoice" + "urography" + "non-fullerene" + "globalgoals" + "complementaryfeeding" + "cops" + "lvcsimulation" + "interatomicpotential" + "tail-dependence" + "computationalphotography" + "opticalvariablesmeasurement" + "theory/philosophicalperspectives" + "safe-haven" + "moderncoexistencetheory" + "agriculturalwaste" + "conjugatedelectronics" + "rfsensor" + "gustatorymmms" + "kinshipanalysis" + "nicheoverlap" + "lightfieldvideo" + "nr3c1" + "globalaml" + "pilottraining" + "hetero-oligomerization"$
   - The size of topics is relatively equal(About 5% of tokens for each). This topic seems to be quite diverse and covers a wide range of fields, from neuroscience ("brain graph") to energy ("crude oil"), education ("school choice"), medical imaging ("urography"), global development ("global goals"), food and nutrition ("complementary feeding"), and more.

2.
   - model with 40 topics and a negative Log Likelihood Bound of -79.6509. One topic in this model is as follows:
   - $"intelligentreflectingsurface" + "cs2agbibr6" + "metasurface" + "non-fullerene" + "smartwindows" + "lead-freedoubleperovskites" + "mitf" + "downlinktraining" + "spheroid" + "wirelessfidelity" + "delayindiagnosis" + "emrs" + "reportingperformance" + "lpcvd" + "ganhemts" + "abscess" + "nhpretreatment" + "greenfinance" + "standardofcare" + "geneticfactors" + "anklebrachialindex" + "immunotoxicology" + "immunostimulation" + "tlr9" + "experimentalmodels" + "hemiallogenic" + "allogenic" + "chemotherapy-inducedperipheralneuropathycipn" + "riskmarker" + "lymphadenectomy"$
   - The size of topics is relatively equal(About 2.5% of tokens for each). The topic itself seems to be primarily focused on advanced materials and wireless communication technologies, with a minor focus on medical and healthcare topics:
     a) **Advanced Materials and Technologies:** The terms "intelligent reflecting surface", "cs2agbibr6", "metasurface", "non-fullerene", "smart windows", and "lead-free double perovskites" all refer to advanced materials or technologies. For instance, "intelligent reflecting surface" refers to a technology that can improve wireless communication by controlling the behavior of electromagnetic waves. "Cs2AgBiBr6" is a type of lead-free double perovskite, which is a promising material for optoelectronic applications due to its high stability and non-toxicity. "Metasurface" is a kind of artificial sheet material that can modulate the behaviors of electromagnetic waves. "Non-fullerene" refers to a type

of acceptor used in organic solar cells as an alternative to fullerenes. “Smart windows” are windows that can change their light transmission properties using advanced materials and technologies.

b) **Wireless Communication:** The term “downlink training” refers to a technique used in frequency division duplexing (FDD) massive MIMO systems to reduce the overhead of the downlink training phase. “Wireless fidelity”, often shortened to Wi-Fi, is a wireless network technology for connecting to the Internet using radio waves.

c) **Medical and Healthcare Topics:** The remaining terms such as “delay in diagnosis”, “emrs” (electronic medical records), “reporting performance”, “abscess”, “nh3 pretreatment”, “standard of care”, “genetic factors”, “ankle brachial index”, “immunotoxicology”, “immunostimulation”, “tlr9” (Toll-like receptor 9), “experimental models”, “hemiallogenic”, “allogenic”, “chemotherapy-induced peripheral neuropathy (CIPN)”, “risk marker”, and “lymphadenectomy” are related to various medical and healthcare topics.

- lastly model with 80 number of topics and the negative Log Likelihood Bound of -161.6086. In this iteration topic modeling converged to a more interesting collection of topics, while a majority of them contain the following:
- $"videostreaming" + "nh3pretreatment" + "nativeoxide" + "lymphadenectomy" + "algorithmicfairness" + "lpcvd" + "ganhemts" + "greenfinance" + "reportingperformance" + "emrs" + "delayindiagnosis" + "abscess" + "standardofcare" + "riskmarker" + "anklebrachialindex" + "immunotoxicology" + "immunostimulation" + "geneticfactors" + "experimentalmodels" + "hemiallogenic" + "allogenic" + "seropositivity" + "tlr9" + "tlr7" + "imiquimod" + "socialdemography" + "policystudies" + "doac" + "rectalcancertreatment" + "3 − yearclinicaloutcome"$
- The topics seem to be primarily focused on advanced materials and technologies, finance and policy, and medical and healthcare.

  a) **Medical and Healthcare Topics:** These terms predominantly relate to medical research, specifically concerning diagnostic processes, treatments, biomarkers, genetic factors, immune responses, and outcomes in clinical settings. For example, "pelvicinsufficiency fractures" and "delay in diagnosis" suggest a focus on orthopedic conditions and diagnostic challenges. "Procollagen (pinp)" and "risk marker" imply research into biomarkers and disease indicators.

  b) **Finance and Policy:** There is evidence suggesting a secondary theme related to finance, policy analysis, and social sciences. "Green finance" and "reporting performance" point to sustainable financial practices and metrics. "Social demography" and "policy studies" indicate a focus on population studies and policy impact, while "algorithmic fairness" could link to ethical considerations in policy and finance algorithms.

  c) **Advanced Materials and Technologies:** There are also terms relevant to material science and engineering, specifically in the context of semiconductor fabrication and surface treatments. The presence of "lpcvd" (Low Pressure Chemical Vapor Deposition), "gan hemts" (Gallium Nitride High Electron Mobility Transistors), "nh3 pretreatment", "native oxide" suggest a focus on advanced manufacturing techniques and materials.

Due to the interdisciplinary nature of this dataset, the results of the large dataset are different from those in subsequent sections, here the aggregation of keywords results in the mixture of interdisciplinary areas. In Figures(5.4,5.5,5.6), some of the top keywords relevant for each

number of topics can be observed. As it was mentioned before and in the dimensionality-reduced results, it is clear that increasing the number of topics would result in more overlap between the generated topics(topics converge to the center of plots).
As a side note5.5, it is observed that increasing the number of training epochs does not have significant effects on the value of log perplexity, and for some cases, it shows an adverse effect in this dataset. Furthermore, this increase will result in higher computational time. After our experiments, value 10 is the preferred number for training epochs in this dataset.

Table 5.5: Number of epochs for different num_topics(lower is better)

	Number of epochs				
num_topics	**10**	**40**	**80**	**100**	**150**
20	-40.3305	-40.3086	-40.3070	-40.3070	-40.3074
40	-79.6509	-79.6503	-79.6507	-79.6508	-79.6511
80	-161.6086	-161.6086	-161.6087	-161.6087	-161.6087

5.1.1 Media and Information Technology publications

The division focuses on both fundamental and applied research [76]. Some of the research in this division includes the Design and validation of a deep evolutionary time visual instrument[100], Use of Embeddings in Visual Analytics [52], and Interactive Visual Exploration of Bibliographic Data [52]. According to the general topics of research, it is expected to observe themes that reflect the division's focus on combining research data with visualization technology to advance science communication and education. After applying the topic modeling algorithm to the group of keywords in each article in recent papers, the model shows different results for the selected number of topics. For this dataset, the LDA algorithm is trained for 40 epochs with a maximum of 5000 iterations for inferring the topic distributions of the text corpus.

1. According to 5.10 The Per-Word Likelihood Bound shows that the gradual increase in the number of topics would increase the performance of the model in fitting the data. The exceptions to this general rule are transitioning during the 76th to 94th period, in other words, fluctuations can be observed for this fitted model.

2. 5.11 shows the Jaccard difference between the discovered topic. In the beginning, there are few similarities between the topics constructed, as the number of topics is increased the clusters become more similar until the selection of 48 topics where two identical clusters are observed. In this process no significant pattern was observed, indicating the random behavior of this cluster distribution.

3. 5.12 shows the distribution of Hellinger Distance between each pair of clusters. Among the top hundred topics, the min value is 0, the max value is 0.8, the boxes represent the middle 50% of the data and the middle line represents the median. In these iterations, the standard deviation fluctuates between 0.02 to 0.2. The data for these iterations can be observed in tabel 5.6.

Table 5.6: The initial iterations for LDA discovered topics according to Hellinger distance.

min	max	mean	median	std_dev	1st_quantile	2nd_quantile	3rd_quantile
0.449372	0.50489	0.476858	0.474877	0.017017	0.465462	0.474877	0.489329
0.451493	0.530287	0.498112	0.500778	0.019001	0.489658	0.500778	0.507923
0.477699	0.537775	0.510459	0.510978	0.015436	0.502343	0.510978	0.523889
0.495263	0.546842	0.521099	0.519968	0.015125	0.507896	0.519968	0.533747
0.494063	0.571074	0.531635	0.533515	0.021282	0.519494	0.533515	0.546298
0.508916	0.577623	0.539602	0.538309	0.017748	0.52978	0.538309	0.548155
0.500242	0.59287	0.548143	0.546082	0.019698	0.5347	0.546082	0.562625
0.505309	0.616472	0.554338	0.55245	0.023232	0.542643	0.55245	0.568173
0.519828	0.619535	0.562212	0.559465	0.022301	0.546922	0.559465	0.575435
0.510721	0.617693	0.565942	0.56668	0.024283	0.549238	0.56668	0.58239
0.50868	0.620293	0.570737	0.567473	0.020618	0.558136	0.567473	0.584003
0.496611	0.642622	0.573193	0.575213	0.029329	0.554613	0.575213	0.592749
0.531901	0.639248	0.581232	0.579362	0.02108	0.565695	0.579362	0.595497
0.470858	0.645582	0.577559	0.583447	0.035239	0.551909	0.583447	0.605471
0.507951	0.672255	0.585789	0.585499	0.029101	0.570972	0.585499	0.599765
0.522699	0.670198	0.58862	0.586285	0.028593	0.568014	0.586285	0.606409

According to the observed measurements and the semantic meaning of observed clusters the two models below are selected:

1.
   - model with 22 number of topics and the negative Log Likelihood Bound of -8.1789. Some topics in this model are as follows:
   - $0.025 \times "industrialoperator" + 0.025 \times "domainshift" + 0.013 \times "researchsoftware" + 0.013 \times "academicsoftware" + 0.013 \times "softwarelicensing"$
   - $0.020 \times "multilayernetworks" + 0.020 \times "graphdrawing" + 0.020 \times "teamwork" + 0.010 \times "educationalsimulations" + 0.010 \times "hinformationtechnologyandsystems"$
   - $0.037 \times "relativisticjet" + 0.019 \times "textvisualization" + 0.019 \times "visualtextanalytics" + 0.019 \times "textmining" + 0.019 \times "naturallanguageprocessing"$
2.
   - model with 6 number of topics and the negative Log Likelihood Bound of -7.7693. Some topics in this model are as follows:
   - $0.046 \times "human-centeredcomputing" + 0.015 \times "visualizationdesignandevaluationmethods" + 0.015 \times "datavisualization" + 0.015 \times "visualizationapplicationdomains" + 0.015 \times "neuralnetworks"$
   - $0.017 \times "ai" + 0.009 \times "levelsofautomation" + 0.009 \times "domainshift" + 0.009 \times "interactivesonification" + 0.006 \times "userevaluation"$
   - $0.014 \times "cancer" + 0.012 \times "relativisticjet" + 0.009 \times "lymphnodes" + 0.009 \times "colon" + 0.006 \times "ccsconcepts"$

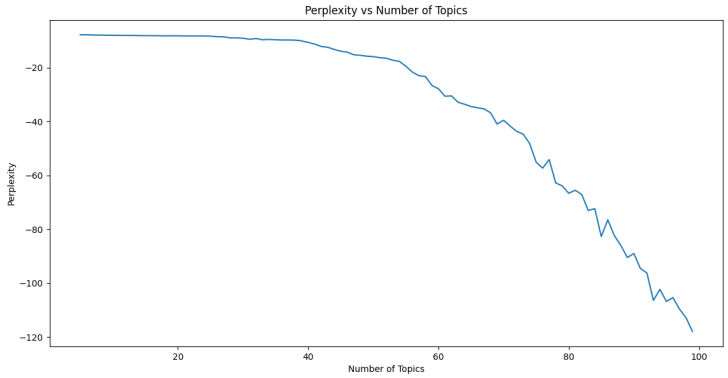


Figure 5.10: Per-Word Log Likelihood Bound (Negative Values) for the number of topics between 5 to 100($4 < n < 99$)

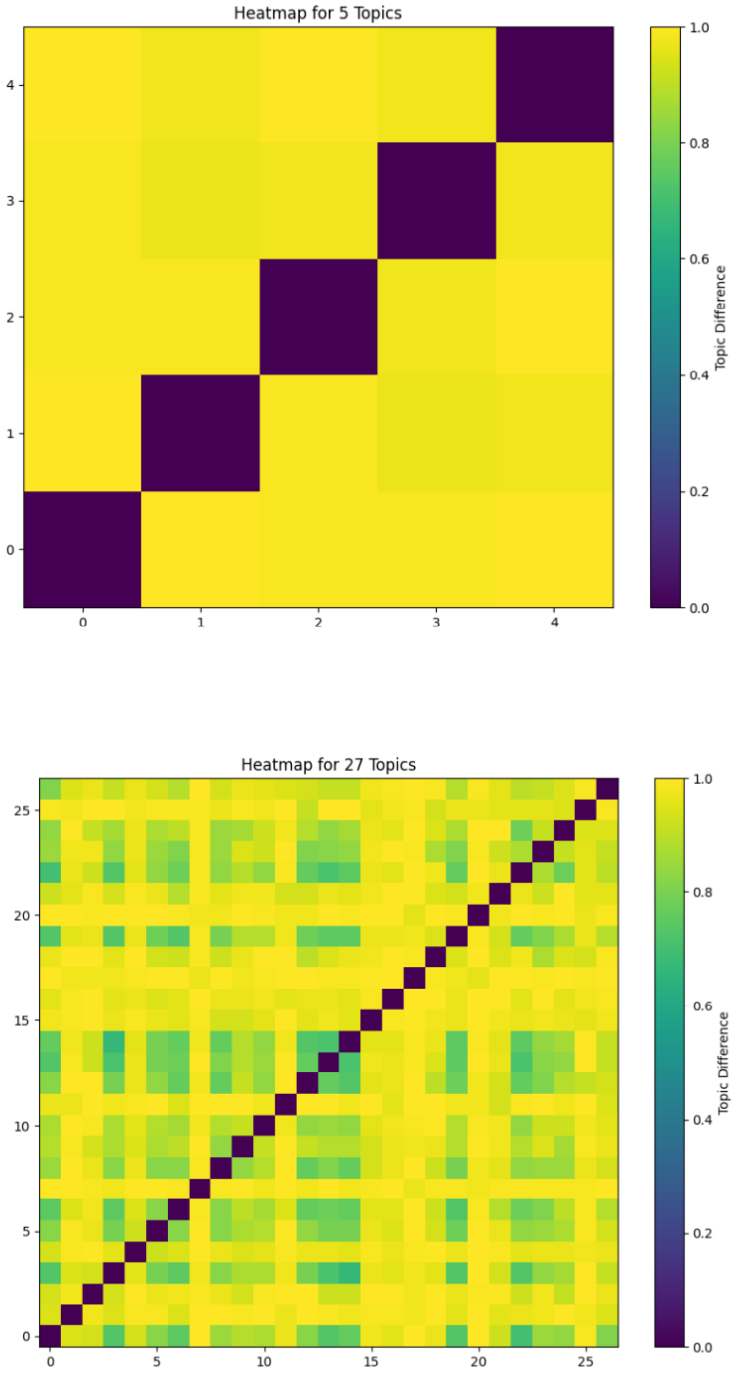


Figure 5.11: Computed topic differences with Jaccard distance measure

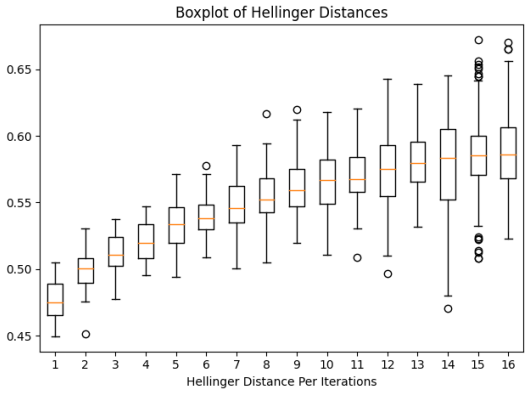


Figure 5.12: Hellinger Distance for the number of topics between 5 to 21($4 < n < 21$)

5.1.2 Computer and Information Science

Department of Computer and Information Science at Linköping University was officially founded in 1983, but some of the papers [44, 28, 57] associated with this institution go back to 1980.
The research topics of this department are majorly dived into Artificial intelligence, "Cognition, interaction and design", Software and computer systems, and Data science [25] with multiple diverse and comprehensive subareas. Some of these broad topics include simulating the return time in a critical-case birth-death process [9] within the realm of statistical mathematics, anomaly detection in healthcare [128] in the field of machine learning, and use of metaheuristic optimization algorithm for allocation of distributed generators in Power Systems and Electrical Engineering domain. Given the general direction of research, it is expected that the active theme of research belongs to the four abovementioned major topics. To explore it further, the LDA algorithm was trained for 40 epochs and a maximum of 5000 iterations for the aim of topic discovery in recent papers.

1. According to 5.13 The Per-Word Likelihood Bound shows that the gradual increase in the number of topics would increase the model's performance in fitting the data. There is no unusual fluctuation and the decrease in value is stable throughout the time period.

2. 5.14 shows the Jaccard difference between the discovered topic. In the beginning, there are few similarities between the topics constructed, as the number of topics is increased the clusters become more similar. By looking at 5.15 it is seen that the selection of 70 topics is the first instance in which two identical clusters are observed. After this iteration, identical clusters tend to be closer spatially and form larger squares successively. In this process no significant pattern was observed, indicating the random behavior of this cluster distribution.

3. 5.16 shows the distribution of Hellinger Distance between each pair of clusters. Among the top hundred topics, the overall minimum value is 0, the max value is 0.7, and the average of the data is 0.5. In these iterations, the standard deviation fluctuates between 0.05 to 0.2. An interesting pattern observed in this dataset is the pattern in the presence of outliers. The majority of outliers occur after iteration 35 of topic modeling and they are always on the lower bound of the Interquartile Range (IQR) 5.17. The data for the iterations between 1 to 16 can be observed in tabel 5.7.

Table 5.7: The initial iterations for LDA discovered topics according to Hellinger distance.

min	max	mean	median	std_dev	1st_quantile	2nd_quantile	3rd_quantile
0.461037	0.489589	0.477067	0.47737	0.007329	0.475987	0.47737	0.480437
0.478541	0.510579	0.493267	0.491721	0.007423	0.489375	0.491721	0.498283
0.492788	0.518481	0.510614	0.51184	0.005901	0.510303	0.51184	0.51344
0.502637	0.532009	0.520329	0.520928	0.006974	0.51603	0.520928	0.524942
0.517753	0.547132	0.530271	0.530135	0.007117	0.525196	0.530135	0.535242
0.514633	0.563617	0.539226	0.542533	0.01177	0.530442	0.542533	0.54781
0.524686	0.582229	0.545411	0.542598	0.012173	0.537491	0.542598	0.550092
0.53715	0.572453	0.55314	0.55323	0.007322	0.548342	0.55323	0.557931
0.509629	0.589113	0.558418	0.559656	0.014928	0.555084	0.559656	0.568183
0.531707	0.58566	0.564493	0.566772	0.011777	0.55836	0.566772	0.572847
0.530637	0.60012	0.568808	0.567504	0.014265	0.559119	0.567504	0.578849
0.524407	0.606158	0.572905	0.574014	0.016004	0.563332	0.574014	0.58448
0.542717	0.609855	0.576923	0.577721	0.014852	0.566223	0.577721	0.586831
0.54421	0.614041	0.582339	0.585105	0.014772	0.57145	0.585105	0.591806
0.54426	0.621746	0.583978	0.583549	0.015175	0.574115	0.583549	0.594728
0.551656	0.616146	0.588873	0.589617	0.011612	0.582186	0.589617	0.597252

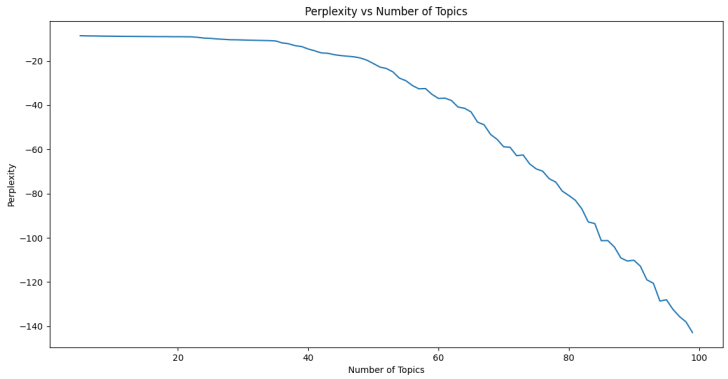


Figure 5.13: Per-Word Log Likelihood Bound (Negative Values) for the number of topics between 5 to 100($4 < n < 99$)

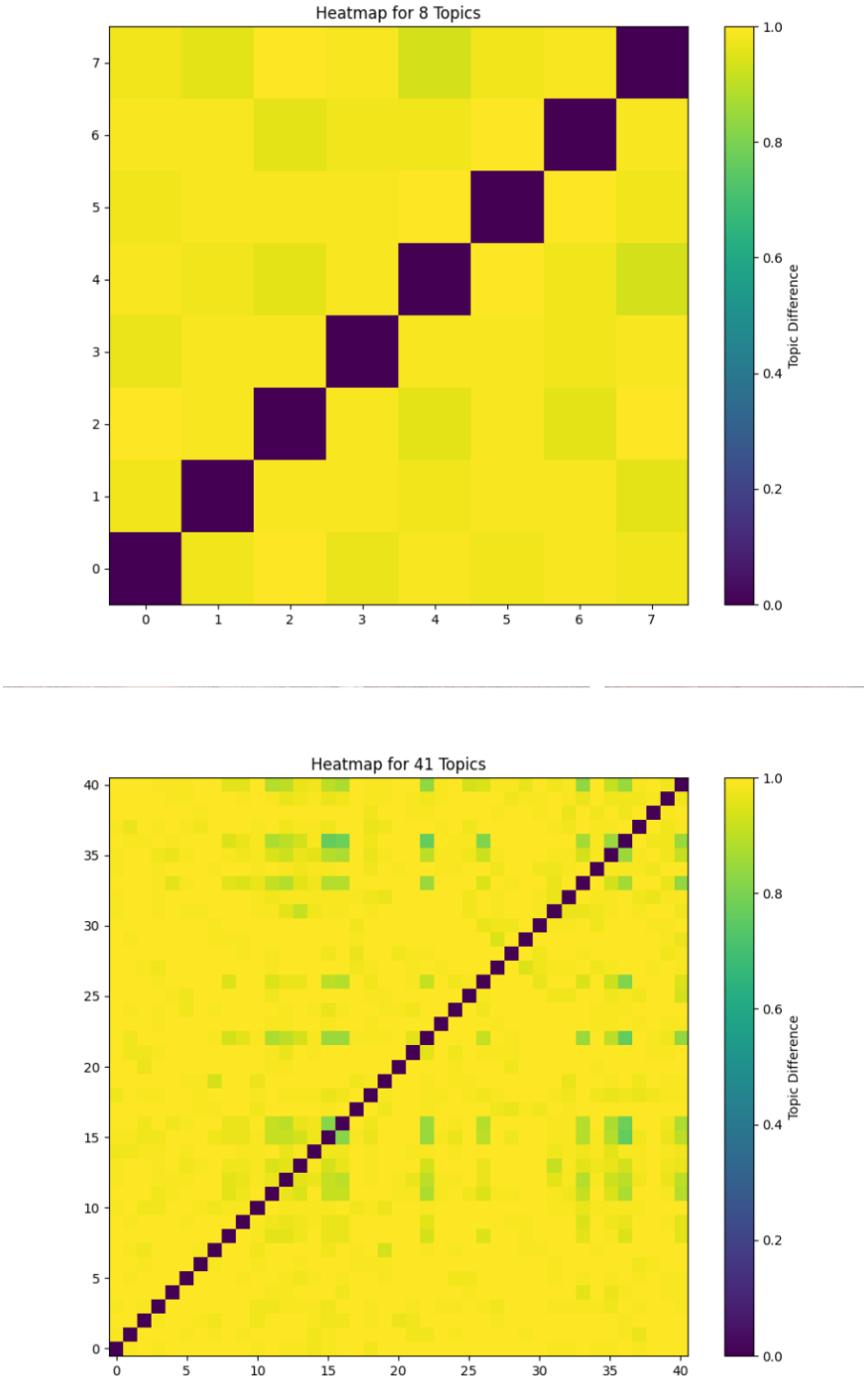


Figure 5.14: topic differences with Jaccard distance measure

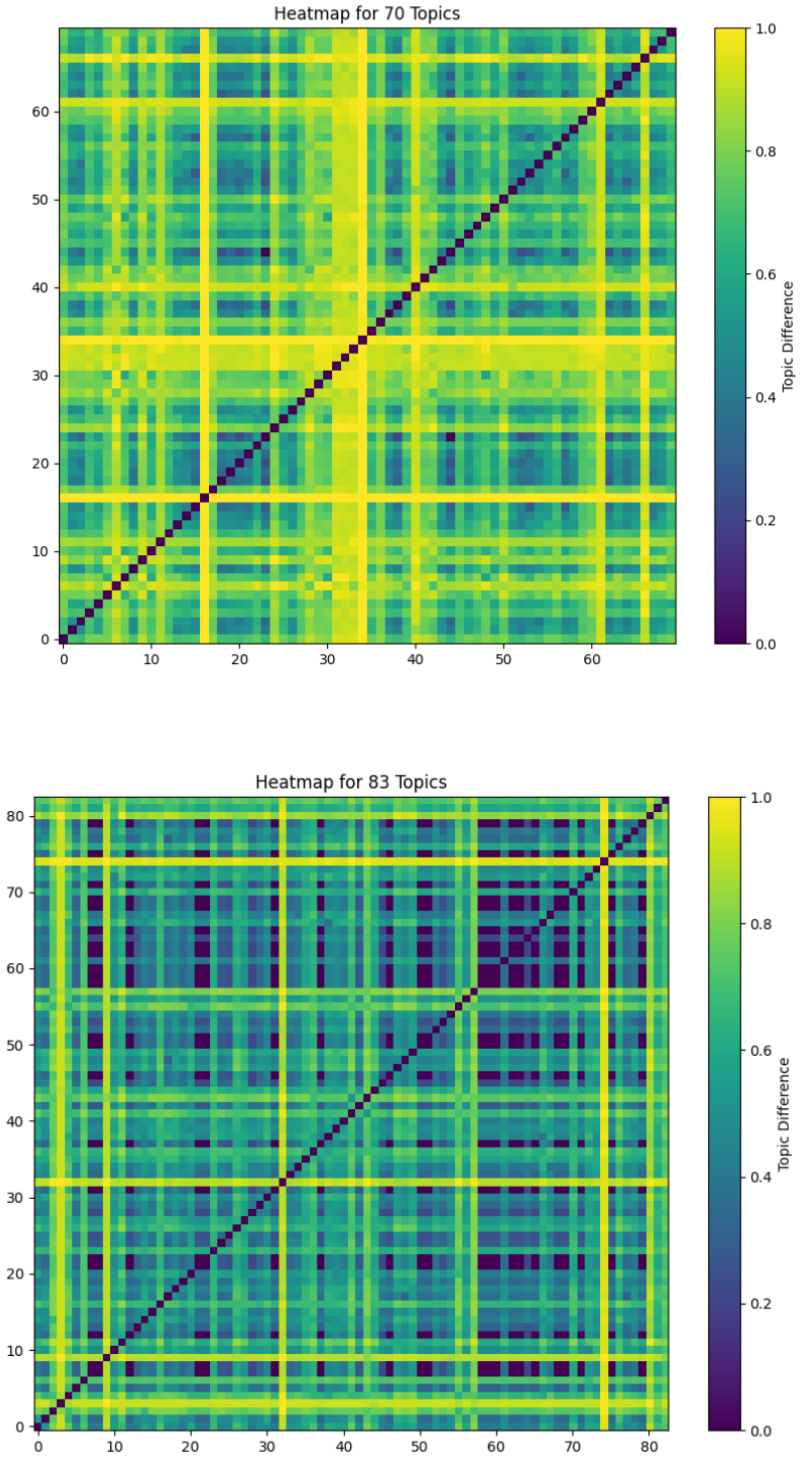


Figure 5.15: Pattern of topic differences with Jaccard distance measure

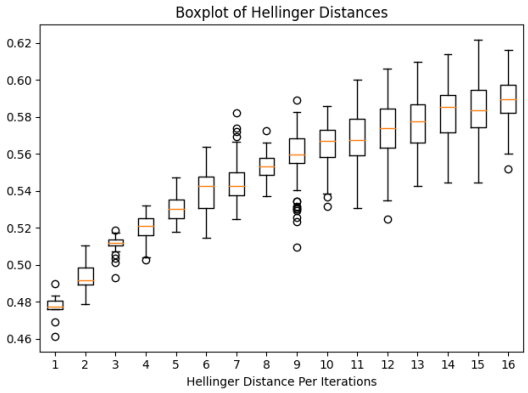


Figure 5.16: Hellinger Distance for the number of topics between 5 to 21($4 < n < 21$)

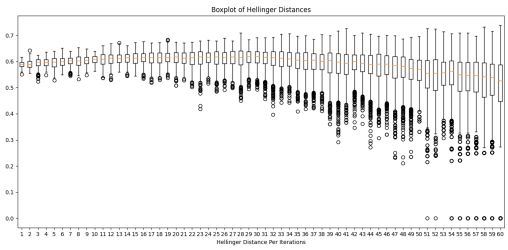


Figure 5.17: Hellinger Distance for the number of topics between 20 to 80($19 < n < 80$)

According to the observed measurements and the semantic meaning of observed clusters the two models below are selected:

1.
   - model with 20 number of topics and the negative Log Likelihood Bound of -9.0392. Some topics in this model are as follows:
   - $0.017 \times "malware" + 0.013 \times "deeplearning" + 0.013 \times "obfuscation" + 0.013 \times "hlso - ochsjukvrd" + 0.009 \times "stringencryption"$
   - $0.019 \times "deeplearning" + 0.012 \times "convolutionalneuralnetworks" + 0.012 \times "modellingforagentbasedsimulation" + 0.008 \times "industrialinternetofthings(iiot)" + 0.008 \times "artificialintelligence(ai)"$
   - $0.016 \times "classicalplanning" + 0.008 \times "delays" + 0.008 \times "deeplearning" + 0.008 \times "primaryschool" + 0.004 \times "decoding"$
2.
   - model with 40 number of topics and the negative Log Likelihood Bound of -14.5865. Some topics in this model are as follows:
   - $0.019 \times "codeobfuscation" + 0.019 \times "disassemblydesynchronization" + 0.019 \times "x86architecture" + 0.019 \times "reverseengineering" + 0.019 \times "moldableparalleltasks"$
   - $0.021 \times "casualexergames" + 0.021 \times "comfortzone" + 0.019 \times "exertion" + 0.011 \times "two - stagegame" + 0.011 \times "gametheory"$
   - $0.025 \times "causalinference" + 0.025 \times "fine - grainedcomplexity" + 0.017 \times "confounding" + 0.017 \times "high - levelparallelprogramming" + 0.017 \times "infinitedomains"$

(5.18, 5.19 The first model specifically successfully managed to cluster many of the relevant keywords together. For instance, topic 10 contains the papers corresponding to reverse engineering(e.g. [74]), and topic 2 mostly contains papers on machine learning and agent-based

systems(e.g. [58]) as well as some reverse engineering papers. It is also noteworthy that some popular keywords such as 'deep learning' have wider distribution among papers, therefore it is contained within 9 topics, the majority being in topics 10 and 2.

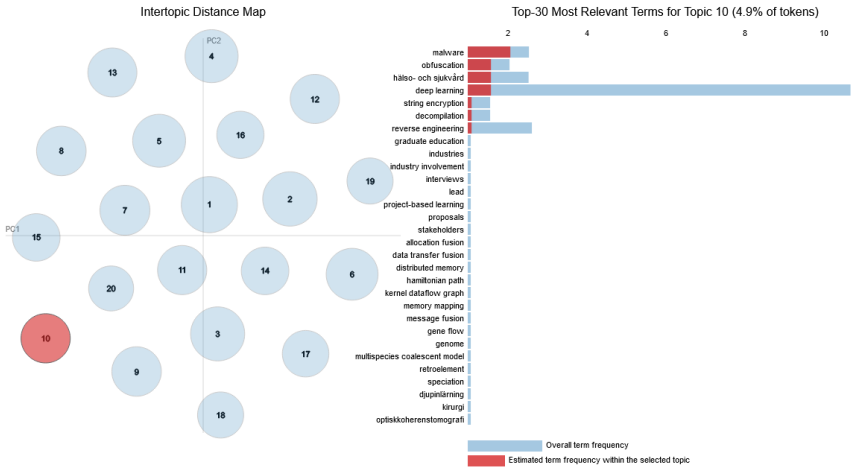


Figure 5.18: LDA with 20 topics - Relevant papers in topics as well as intertopic Distance MAP

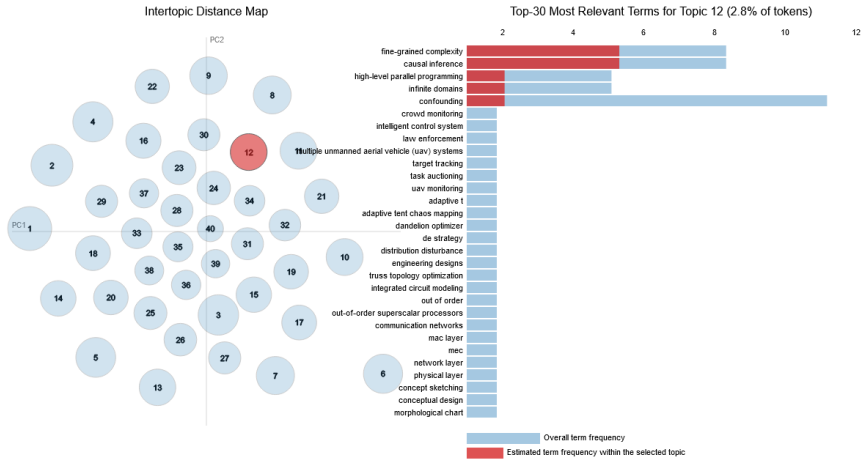


Figure 5.19: LDA with 40 topics - Relevant papers in topics as well as intertopic Distance MAP

5.2 Analysis of influence

The main aim of this section is to perform a comprehensive analysis of the relationship between keywords.

5.2.1 Media and Information Technology publications

This subset, which consists of 1055 published articles, spans from the year 2000 to the present. The dataset follows the general structure of the large dataset, overall This sub dataset is particularly well-suited for our research due to its specificity to the field of scientometrics.

5.2.1.1 Vector Autoregression

In this section VAR method was performed on the keywords with most occurrences. Out of the 2205 keywords, the top 220 pair combinations were selected which resulted in 137 unique keywords for analysis.
All variables in a VAR enter the model in the same way: each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term. As a result of VAR modeling on 74 stationary time series, 2700 equations were generated.
In these models, the AIC measure was chosen to determine the optimal number of lags to include in the model. Some of the interesting results in this batch are mentioned below:

1. For variables calibration and active illumination, the number of observations is 19 and the smallest AIC score belongs to lag 3
   For the 'active illumination' equation:

$$\begin{aligned} AI_t = &-2.00 \times AI_{t-2} + 2.00 \times C_{t-2} \\ &- 2.00 \times AI_{t-3} + 1.00 \times C_{t-3} + \epsilon_{AI} \end{aligned}$$

   and For the 'calibration' equation:

$$\begin{aligned} C_t = &-2.209 \times AI_{t-2} + 1.923 \times C_{t-2} \\ &- 1.637 \times AI_{t-3} + \epsilon_C \end{aligned}$$

2. For variables medical visualization and molecular life science, the number of observations is 18, and the smallest AIC score belongs to lag 4
   For the 'medical visualization' equation:

$$\begin{aligned} MV_t = &-0.0523 \times MV_{t-1} + 1.298 \times MLS_{t-2} + 0.562 \times MLS_{t-3} - 0.617 MV_{t-4} \\ &+ 0.000174 MLS_{t-4} + \epsilon_{MV} \end{aligned}$$

   For the 'molecular life science' equation:

$$\begin{aligned} MLS_t = &1.236 \times MV_{t-2} \\ &- 0.741 \times MV_{t-3} - 0.626 \times MLS_{t-4} + \epsilon_{MLS} \end{aligned}$$

3. For variables active illumination and geovisual analytics, the number of observations is 17, and the smallest AIC score belongs to lag 5
   For the 'active illumination' equation:

$$\begin{aligned} AI_t = &-2.00 \times AI_{t-1} + 2.00 \times GA_{t-2} \\ &- 3.00 \times AI_{t-3} + 1.00 \times GA_{t-3} - 3.00 \times AI_{t-4} - 2.00 \times AI_{t-5} + \epsilon_{AI} \end{aligned}$$

   For the 'geovisual analytics' equation:

$$\begin{aligned} GA_t = &\, 6.510 \times AI_{t-1} + 1.857 \times GA_{t-1} - 2.183 \times AI_{t-2} \\ &- 3.857 \times GA_{t-2} + 5.020 \times AI_{t-3} - 2.447 \times GA_{t-3} \\ &+ 5.089 \times AI_{t-4} + 4.858 \times AI_{t-5} + \epsilon_{GA} \end{aligned}$$

4. For variables shading and shadowing, the number of observations is 17 and the smallest AIC score belongs to lag 5, additionally 5.20shows the next 20 timesteps of these two

variables, it can be seen that the forecasted values of these exhibit a certain pattern. Both variables appear to have a slight upward trend, indicating an increase in intensity and then later experiencing a slight decrease. After 14 points of timesteps, they seem to remain relatively stable, with minor fluctuations around their mean value.
For the 'shading' equation:

$$\begin{aligned}\text{shading}_t = 0.285714 \\ &+ 0.107 \times \text{shading}_{t-1} \\ &+ 0.250 \times \text{shadowing}_{t-1} \\ &+ 0.571 \times \text{shading}_{t-2} \\ &- 0.392 \times \text{shadowing}_{t-2} \\ &+ 1.571 \times \text{shading}_{t-3} \\ &- 1.857 \times \text{shadowing}_{t-3} \\ &- 0.357 \times \text{shading}_{t-4} \\ &+ 0.071 \times \text{shadowing}_{t-4} \\ &- 0.464 \times \text{shading}_{t-5} \\ &+ 0.179 \times \text{shadowing}_{t-5} + \epsilon_{\text{shading}}\end{aligned}$$

and for the 'shadowing' equation:

$$\begin{aligned}\text{shadowing}_t = 0.286 \\ &+ 0.607 \times \text{shading}_{t-1} \\ &- 0.250 \times \text{shadowing}_{t-1} \\ &+ 0.571 \times \text{shading}_{t-2} \\ &- 0.893 \times \text{shadowing}_{t-2} \\ &+ 1.571 \times \text{shading}_{t-3} \\ &- 1.857 \times \text{shadowing}_{t-3} \\ &- 0.357 \times \text{shading}_{t-4} \\ &+ 0.071 \times \text{shadowing}_{t-4} \\ &+ 0.036 \times \text{shading}_{t-5} \\ &- 0.321 \times \text{shadowing}_{t-5} + \epsilon_{\text{shadowing}}\end{aligned}$$

Among the first three variable pairs, coefficients that satisfy the p-value of t-statistics, in other words, those that are statistically significant are selected.
It is noted that for variables that correspond to keyword co-occurrence in this data, they would mostly have a correlation of 1, or with some time lag, the future correlation would change to 1. Among the 74 stationary variables tested using VAR, 45 combinations are among the $co-oocurred$ keywords and 26 of them correlate to 1.
Out of the rest of the keyword combinations, 26 pairs are completely identical(therefore correlate to 1). Only the following combinations have the property of uniqueness and non-existing in keyword co-occurrence:

- The pair **('layered component architecture', 'multimedia information systems')**
- The pair **('multimedia information systems', 'web-enabled visualization toolkit')**
- The pair **('layered component architecture', 'multimedia information systems')**
- The pair **('multimedia information systems', 'web-enabled visualization toolkit')**

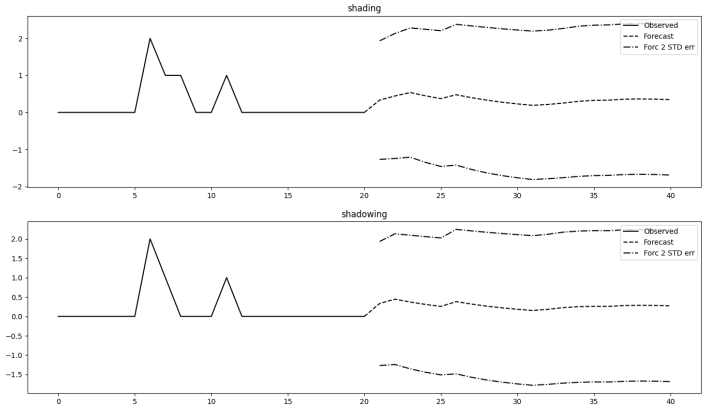


Figure 5.20: 20 Future timesteps for 'shading', 'shadowing' keywords

5.2.1.2 Granger Causality

The Granger causality test can be used to investigate the causal relationships between the variables in the VAR model. For each equation in a VAR system, using the Granger causality test, we can check the hypotheses that each of the keywords in the data can Granger-cause the target variable in that equation. In this section granger causality was performed on the most frequent keywords in this dataset. Table 5.8,5.9 show the results of the Granger causality test on 30 keywords with a time lag of 5.
from the first table we have :

1. **Scientific Visualization and Visualization**: The p-values for lags 1 to 4 are less than 0.05, suggesting that there is a strong Granger causality from "visualization" to "scientific visualization" for these lags. This means changes in "visualization" can predict changes in "scientific visualization".

2. **Multimedia Information Systems and User Interfaces and Human Computer Interaction**: The p-values for lag 1, and lag 2 are less than 0.05, indicating that "user interfaces and human computer interaction" Granger causes "multimedia information systems" at these lags.

3. **Computers and Society and Multimedia Information Systems**: The p-values for lag 1 to lag 5 are all less than 0.05, suggesting a strong Granger causality from "multimedia information systems" to "computers and society". This means that past values of "multimedia information systems" can help predict future values of "computers and society".

4. **Authentic Learning and Technology Education**: The p-values for lag 1 and lag 4 are less than 0.05, and for lag 5 it's 0, indicating that "technology education" Granger causes "authentic learning" at these lags.

5. **Graininess and Multi-channel Printing**: The p-values for all lags are 0, which suggests a very strong Granger causality from "multi-channel printing" to "graininess". This means that changes in "multi-channel printing" can predict changes in "graininess".

and from the second table we have :

1. **"computers and society" and "computer science"**: The p-values are less than 0.05 for all lags, suggesting that the occurrence of "computer science" can significantly predict the occurrence of "computers and society" at all tested lags.

2. **"multimedia information systems" and "computer science"**: The p-values are less than 0.05 for lags 1 and 2, suggesting that "computer science" can predict "multimedia information systems" at these lags. However, the p-values are greater than 0.05 for lags 3 to 5, indicating no significant prediction at these lags.

3. **"shadowing" and "shading"**: The p-values are greater than 0.05 for all lags, suggesting that "shading" does not significantly predict "shadowing" at any of the tested lags.

4. **"multimedia systems" and "multimedia information systems"**: The p-values are less than 0.05 for all lags, suggesting that "multimedia information systems" can significantly predict "multimedia systems" at all tested lags.

5. **"magnetic resonance imaging" and "computational fluid dynamics"**: The p-values are less than 0.05 for lags 1 to 4, suggesting that "computational fluid dynamics" can significantly predict "magnetic resonance imaging" at these lags. However, the p-value is greater than 0.05 for lag 5, indicating no significant prediction at this lag.

Table 5.8: Granger Causality test - The p-values are based on F-test for media and information science [Keyword 2 Granger causes Keyword 1]

Keywords 1	Keywords 2	lag 1	lag 2	lag 3	lag 4	lag 5
shading	shadowing	0.3418	0.5468	0.4045	0.5283	0.7612
computer science	computers and society	0.9591	0.7935	0.1358	0.0185	0.0262
computer science	multimedia information systems	0.047	0.8035	0.8853	0.3633	0.2055
scientific visualization	visualization	0.0001	0.0014	0.0116	0.0262	0.0724
ellipses	plane estimation	0.1653	0.395	0.5686	0.7603	0.4973
multimedia systems	user interfaces and human computer interaction	0.9547	0.2451	0.4317	0.202	0.3999
technology education	technology teachers	0.6752	0.3712	0.1615	0.2982	0.0063
evolution	threshold concepts	0.6254	0.9455	0.9631	0.9339	0.9567
life science	visualization	0.498	0.7807	0.7969	0.9103	0.9782
multimedia information systems	user interfaces and human computer interaction	0.0008	0.0362	0.0573	0.1736	0.4227
computers and society	multimedia information systems	0.0046	0.0001	0.0006	0.0043	0.0384
active illumination	plane estimation	0.0187	0.0415	0.0666	0.1439	0.3278
cancer	pathology	0.5787	0.5448	0.1746	0.168	0.1397
computer science	user interfaces and human computer interaction	0.9605	0.7823	0.9201	0.3898	0.6796
active illumination	ellipses	0.9304	0.5835	0.2592	0.1736	0.4557
visualization	visualization design and evaluation methods	0.4791	0.0044	0.1868	0.3453	0.182
evolution	visualization	0.2103	0.4965	0.6947	0.8785	0.8838
multimedia information systems	multimedia systems	0.0008	0.0363	0.1294	0.4591	0.3962
computational fluid dynamics	magnetic resonance imaging	0.194	0.1184	0.0016	0.0181	0.127
ellipses	torchlight	0.1402	0.4623	0.7032	0.5285	0.3973
active illumination	torchlight	0.4959	0.6874	0.7661	0.8469	0.2347
computer science	multimedia systems	0.9562	0.7826	0.9225	0.9479	0.9675
graininess	multi-channel printing	0	0	0	0	0
plane estimation	torchlight	0.9082	0.4419	0.5674	0.7037	0.4766
authentic learning	technology education	0.0103	0.0738	0.1842	0.0248	0
information visualization	visualization	0.2719	0.4859	0.6497	0.3292	0.4855
computers and society	user interfaces and human computer interaction	0.5087	0.9034	0.5465	0.5462	0.7522
geovisual analytics	storytelling	0.3423	0.5042	0.7371	0.8986	0.2534
color prediction	dot gain	0.3315	0.5564	0.0312	0.244	0.5647
computers and society	multimedia systems	0.1613	0.279	0.0912	0.1107	0.1712

Table 5.9: Granger Causality test in reverse- The p-values are based on F-test for media and information science [Keyword 2 Granger causes Keyword 1]

Keywords 1	Keywords 2	lag 1	lag 2	lag 3	lag 4	lag 5
shadowing	shading	0.7767	0.9359	0.5611	0.6579	0.8777
computers and society	computer science	0.0101	0	0.0002	0.0028	0.0176
multimedia information systems	computer science	0.0207	0.0366	0.1315	0.4216	0.3356
visualization	scientific visualization	0.0156	0.0176	0.0388	0.1869	0.1067
plane estimation	ellipses	0.7671	0.8766	0.9506	0.9825	0.9876
user interfaces and human computer interaction	multimedia systems	0.9898	0.7251	0.8292	0.7266	0.8627
technology teachers	technology education	0.6215	0.8262	0.9311	0.9827	0.3084
threshold concepts	evolution	0.8242	0.9393	0.9019	0.9722	0.3737
visualization	life science	0.7351	0.8437	0.795	0.5064	0.2217
user interfaces and human computer interaction	multimedia information systems	0.0046	0.0001	0.001	0.0143	0.0776
multimedia information systems	computers and society	0.0008	0.0363	0.1329	0.2066	0.1572
plane estimation	active illumination	0.2913	0.435	0.6768	0.7885	0.6719
pathology	cancer	0.6466	0.1016	0.2146	0.4492	0.4815
user interfaces and human computer interaction	computer science	0.0098	0	0.0006	0.0049	0.0404
ellipses	active illumination	0.9806	0.9795	0.8869	0.9533	0.9597
visualization design and evaluation methods	visualization	0.0162	0.0363	0.0735	0.0637	0.2813
visualization	evolution	0.5494	0.8776	0.634	0.8995	0.6723
multimedia systems	multimedia information systems	0.0046	0.0001	0.0009	0.0132	0.0118
magnetic resonance imaging	computational fluid dynamics	0.0004	0.0083	0.0009	0.0093	0.0828
torchlight	ellipses	0.4749	0.5109	0.046	0.0514	0.1091
torchlight	active illumination	0.041	0.0289	0.0626	0.1819	0.1457
multimedia systems	computer science	0.0104	0	0.0006	0.0081	0.0588
multi-channel printing	graininess	0.0061	0.0846	0.2681	0.223	0.1052
torchlight	plane estimation	0.8777	0.8797	0.9828	0.989	0.9013
technology education	authentic learning	0.5282	0.3937	0.0167	0	0.002
visualization	information visualization	0.5242	0.4986	0.7994	0.9662	0.9453
user interfaces and human computer interaction	computers and society	0.3268	0.5344	0.2774	0.5022	0.0573
storytelling	geovisual analytics	0.3788	0.7735	0.8795	0.9634	0
dot gain	color prediction	0.5657	0.0022	0.0034	0.0232	0.0937
multimedia systems	computers and society	0.0344	0.0784	0.1183	0.1492	0.3688

5.2.1.3 Bayesian Network Analysis

The structure that is defined by keyword relationships and the structure learned by the structural learning algorithm for this dataset show significant differences. For both cases, the structure contains 64 nodes of the relevant keywords. The custom-defined network contains 80 directed edges, representing the connection between the keywords and the structure learned by the MIIC algorithm contains 64 nodes.
Table 5.10, 5.11 show the marginal probabilities of two networks. In a Bayesian network, marginal probabilities represent the likelihood of each state of a variable, independent of the states of other variables. Each keyword is a variable with two possible states: 0 (the keyword does not occur) and 1 (the keyword does occur).
These probabilities help us understand the prevalence of each keyword in the dataset. For example, 'Air Traffic Control', 'Deep Learning', and 'Sonification' have the highest probabilities of not occurring (around 0.6956) compared to the other top keywords, suggesting they are less common in this dataset. As the most frequent keyword, 'visualization' has the the closest probabilities for occurring and not occurring (0.5004 and 0.4996 respectively), indicating it is roughly equally likely to occur as not to occur in this dataset.
On the other hand, in the case of the learned Bayesian network, we can observe that the probabilities of occurrence (state 1) for all keywords are significantly lower in the learned network. This suggests that the MIIC algorithm has learned a network where these keywords are less likely to occur independently.
For example, the probability of 'Visualization' occurring in the designed network was approximately 0.5, but in the learned network, it is approximately 0.07. This could suggest that the occurrence of 'Visualization' is more dependent on other variables in the learned network.

The figures 5.21 and 5.22 show the two Bayesian networks visualized by directed acyclic graphs. Each node in the network represents a keyword, and the edges (lines) between nodes represent dependencies or relationships between the keywords. In the first structure the nodes corresponding to 'augmented reality', 'evolution', and 'deep learning' are the ancestral nodes, while 'machine learning', 'visual analytics', and 'visualizations' are crucial hubs

Table 5.10: Marginal Probabilities for top keywords in Designed Network

	0	1
Volume Rendering	0.5512	0.4488
Information Visualization	0.6358	0.3642
PIC Simulation	0.6370	0.3630
Scientific Visualization	0.6369	0.3631
Machine Learning	0.5465	0.4535
Air Traffic Control	0.6956	0.3044
Deep Learning	0.6959	0.3041
Sonification	0.6956	0.3044
Visual Analytics	0.5331	0.4669
Visualization	0.5004	0.4996

Table 5.11: Marginal Probabilities for Top Keywords in Learned Network

	0	1
Volume Rendering	0.9805	0.0195
Information Visualization	0.9792	0.0208
PIC Simulation	0.9820	0.0180
Scientific Visualization	0.9834	0.0166
Machine Learning	0.9790	0.0210
Air Traffic Control	0.9834	0.0166
Deep Learning	0.9842	0.0158
Sonification	0.9834	0.0166
Visual Analytics	0.9752	0.0248
Visualization	0.9275	0.0725

connecting to most other nodes within the graph. In the second structure, the number of ancestral nodes is fewer and the common ancestors are 'authentic learning', 'privacy', and 'graininess'. Additionally, the 'label placement' keyword has a different set of relationships, due to the connection to 'haptics'.
The designed network can be divided into two distinct sections:

- **Left Section**: This section appears densely packed with nodes and edges, indicating a complex, non-linear structure. The high degree of interconnection suggests that the keywords in this section are highly interdependent. In other words, the presence or absence of some of the keywords(i.e. visualization) is likely to influence the presence or absence of many other keywords in this section.
- **Right Section**: This section shows a more structured, hierarchical pattern of connections among nodes. The nodes are connected in series with limited cross-connections, suggesting a more linear dependency structure. This means that the presence or absence of a keyword is mainly influenced by one or two other keywords, rather than a large group of keywords.

On the other hand, the structure of the learned networks is organized into distinct layers, indicating different levels for this network. The high degree of interconnection in the left section of the designed network suggests that many keywords are highly interdependent, in contrast, the layered structure of the learned network suggests a more hierarchical dependency structure, where the presence or absence of a keyword is mainly influenced by the keywords in the layer above it.

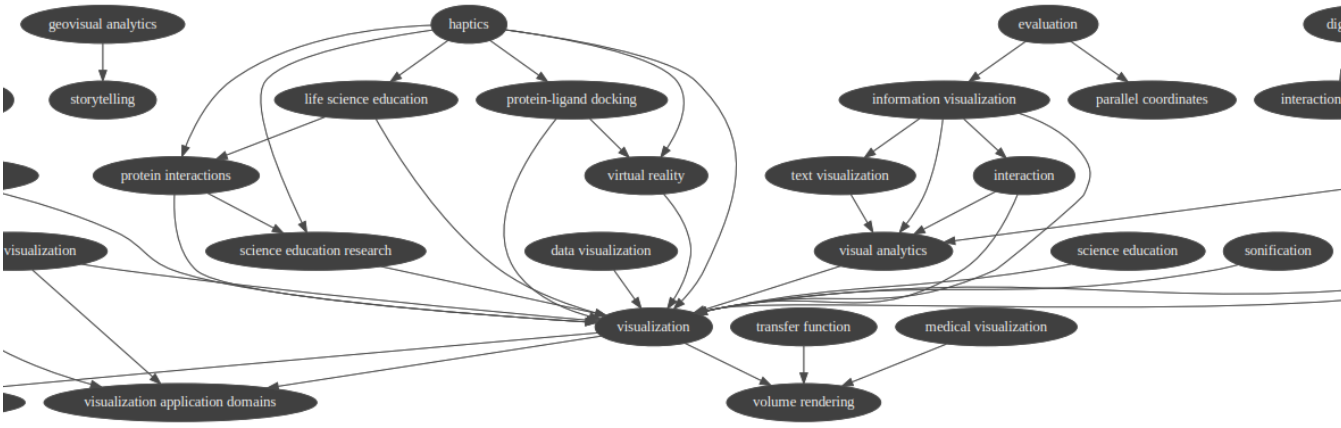


Figure 5.21: Designed structure based on keyword co-occurrence

Figure 5.22: Structure learned by MIIC algorithm based on BDeu score

For a more comprehensive visualization refer to 8.2.
To understand these networks better, the conditional probability of occurrence of the keywords, given some evidence needs to be calculated. Table 5.12 and 5.13 shows the conditional probabilities of the occurrence of certain keywords (Node Name) given the presence of other keywords (Evidence) in your dataset.

Table 5.12: Conditional probabilities of keyword nodes given evidence for designed network

Node Name	Evidence	P(0)	P(1)
Information Visualization	Visual Analytics	0.6103	0.3897
Deep Learning	Machine Learning	0.6649	0.3351
Visual Analytics	Visualization	0.5327	0.4673
Collisionless Plasma	PIC Simulation	0.5840	0.4160
Scientific Visualization	Visualization	0.6366	0.3634
Air Traffic Control	Automation	0.5795	0.4205
Human-Centered Computing	Visualization	0.6971	0.3029
Visual Analytics	Information Visualization	0.5004	0.4996
Machine Learning	Deep Learning	0.5004	0.4996
Visualization	Visual Analytics	0.5000	0.5000
PIC Simulation	Collisionless Plasma	0.4980	0.5020
Visualization	Scientific Visualization	0.5000	0.5000
Automation	Air Traffic Control	0.5015	0.4985
Visualization	Human-Centered Computing	0.5000	0.5000

Table 5.13: Conditional probabilities of keyword nodes given evidence for learned network

Node Name	Evidence	P(0)	P(1)
Information Visualization	Visual Analytics	0.8712	0.1288
Deep Learning	Machine Learning	0.7456	0.2544
Visual Analytics	Visualization	0.8537	0.1463
Collisionless Plasma	PIC Simulation	0.6224	0.3776
Scientific Visualization	Visualization	0.9834	0.0166
Air Traffic Control	Automation	0.2826	0.7174
Human-Centered Computing	Visualization	0.9246	0.0754
Visual Analytics	Information Visualization	0.8465	0.1535
Machine Learning	Deep Learning	0.6628	0.3372
Visualization	Visual Analytics	0.5727	0.4273
PIC Simulation	Collisionless Plasma	0.1190	0.8810
Visualization	Scientific Visualization	0.9274	0.0726
Automation	Air Traffic Control	0.6333	0.3667
Visualization	Human-Centered Computing	0.4856	0.5144

A series of do-calculus is performed to analyze the causal impact in both the designed and the learned networks. The results for the top 5 and their reverse form are as follows:

1.

$$P(\text{visual analytics} \mid \text{do}(\text{visualization})) = \sum_{\text{science education}} P(\text{science education} \mid \text{visualization}) \cdot \left(\sum_{\text{visualization}'} P(\text{visual analytics} \mid \text{science education}, \text{visualization}') \cdot P(\text{visualization}') \right)$$

This equation suggests that the presence of the keyword “visualization” has an influence on the keyword “visual analytics” through the keyword “science education”. This is known as a frontdoor path. In other words, “visualization” affects “science education”, which in turn affects “visual analytics”.

2.

$$P(\text{haptics} \mid \text{do}(\text{visualization})) = \sum_{\text{science education}} P\left(\text{science education} \mid \text{visualization}\right) \cdot \left(\sum_{\text{visualization}'} P\left(\text{haptics} \mid \text{science education}, \text{visualization}'\right) \cdot P\left(\text{visualization}'\right.\right.$$

The second equation is similar to the first one, but here the keyword "haptics" is being influenced by "visualization" through "science education".

3.

$$P(\text{human-centered computing} \mid \text{do}(\text{visualization})) = \sum_{\text{science education}} P\left(\text{science education} \mid \text{visualization}\right) \cdot \left(\sum_{\text{visualization}'} P\left(\text{human-centered computing} \mid \text{science education}, \text{visualization}'\right) \cdot P\left(\text{visualization}'\right) \right)$$

The third equation is also similar, but the keyword "human-centered computing" is being influenced by "visualization" through "science education".

4.

$$P(\text{collisionless plasma} \mid \text{do}(\text{pic simulation})) = P\left(\text{collisionless plasma}\right)$$

This equation suggests that the presence of the keyword "pic simulation" does not have any influence on the keyword "collisionless plasma". This is indicated by the fact that the probability of "collisionless plasma" given the do-operator applied to "pic simulation" is equal to the marginal probability of "collisionless plasma".

5.

$$P(\text{air traffic control} \mid \text{do}(\text{automation})) = P\left(\text{air traffic control}\right)$$

Similar to the fourth equation, this equation suggests that the presence of the keyword "automation" does not have any influence on the keyword "air traffic control".

reversing the causal effect relationship:

1. In this case, the backdoor variables are 'information visualization', 'machine learning', and 'interaction'. These variables are likely confounders that affect both "visual analytics" (the treatment) and "visualization" (the outcome).
2. The reverse form for the causal effect of "haptics" on "visualization" involves summing over all possible values of a large set of variables including 'data visualization', 'deep learning', 'digital design', 'digital pathology', 'evaluation', 'evolution', 'human-centered computing', 'information visualization', 'interaction', 'interaction design', 'life science education', 'machine learning', 'protein interactions', 'protein-ligand docking', 'science education', 'science education research', 'scientific visualization', 'sonification', 'text visualization', 'virtual reality', 'visual analytics'. The presence of these variables in the do-calculus computation suggests that "haptics" influence "visualization" through a complex network of causal pathways involving these variables, but the impact is not impressive enough to justify such complex computation(50%)

3. Similar to the previous equation, the effect of "human-centered computing" on "visualization" involves a large set of variables including 'data visualization', 'deep learning', 'digital design', 'digital pathology', 'evaluation', 'evolution', 'haptics', 'information visualization', 'interaction', 'interaction design', 'life science education', 'machine learning', 'protein interactions', 'protein-ligand docking', 'science education', 'science education research', 'scientific visualization', 'sonification', 'text visualization', 'virtual reality', 'visual analytics'.

4. $$P(picsimulation \mid \text{do}(collisionlessplasma)) = P\left(picsimulation \mid collisionlessplasma\right)$$

   In this case, there is a direct effect of "collisionless plasma" on "pic simulation". Also, the presence of no other variables in the do-calculus computation suggests that "collisionless plasma" influences "pic simulation" directly, and there are no other confounding variables or front door paths in the model.

5. $$P(automation \mid \text{do}(airtrafficcontrol)) = P\left(automation \mid airtrafficcontrol\right)$$

   for these two keywords, there is also a direct influence of "air traffic control" on "automation" without any additional keywords intervening.

By looking at the learned network:

1. $$P(\text{visual analytics} \mid \text{do}(\text{visualization})) = \sum_{\text{digital design}} P(\text{digital design} \mid \text{visualization}) \cdot \left(\sum_{\text{visualization}'} P(\text{visual analytics} \mid \text{visualization}') \cdot P(\text{visualization}') \right)$$

   there exists a frontdoor variable 'digital design'. This variable is a mediator that is affected by "visualization" (the treatment) and in turn affects "visual analytics" (the outcome).

2. $$P(haptics \mid \text{do}(visualization)) = P\left(haptics \mid visualization\right)$$

   This expression suggests that "visualization" influences "haptics" directly.

3. $$P(\text{human-centered computing} \mid \text{do}(\text{visualization})) = \sum_{\text{information visualization}} P(\text{human-centered computing} \mid \text{information visualization}, \text{visualization}) \cdot P(\text{information visualization})$$

   using the backdoor adjustment formula, the backdoor variable is 'information visualization'. This also suggests that while "visualization" influences "human-centered computing", 'information visualization' is a confounder in this case.

4.

$$P(\text{collisionless plasma} \mid \text{do}(\text{pic simulation})) = \sum_{\text{shock}} P(\text{shock} \mid \text{pic simulation}) \cdot \left(\sum_{\text{pic simulation}'} P(\text{collisionless plasma} \mid \text{pic simulation}') \cdot P(\text{pic simulation}') \right)$$

The frontdoor variable 'shock' is the meditator for the relationship of pic simulation" (the treatment) and "collisionless plasma" (the outcome).

5.

$$P(\text{air traffic control} \mid \text{do}(\text{automation})) = \sum_{\text{digital design}} P(\text{digital design} \mid \text{automation}) \cdot \left(\sum_{\text{automation}'} P(\text{air traffic control} \mid \text{automation}') \cdot P(\text{automation}') \right)$$

Similar to the first case, 'digital design' is a frontdoor variable that suggests that "automation" influences "air traffic control" not only directly, but also indirectly through 'digital design'.

when considering the reverse case for these variables:

1.

$$P(\textit{visualization} \mid \text{do}(\textit{visualanalytics})) = P\left(\textit{visualization} \mid \textit{visualanalytics}\right)$$

There is a direct effect of 'visual analytics' on 'visualization'.

2.

$$P(\text{visualization} \mid \text{do}(\text{haptics})) = \sum_{\text{digital design}} P(\text{digital design} \mid \text{haptics}) \cdot \left(\sum_{\text{haptics}'} P(\text{visualization} \mid \text{haptics}') \cdot P(\text{haptics}') \right)$$

The frontdoor variable is 'digital design' and it is a mediator that is affected by "haptics" (the treatment) and in turn affects "visualization" (the outcome).

3.

$$P(\text{visualization} \mid \text{do}(\text{human-centered computing})) = \sum_{\text{visualization design and evaluation methods}} P(\text{visualization design and evaluation methods} \mid \text{human-centered compu} \left(\sum_{\text{human-centered computing}'} P(\text{visualization} \mid \text{human-centered computing}') \cdot P(\text{human-centered computing}') \right)$$

The keyword 'visualization design and evaluation methods' is the frontdoor variable for the effect of "human-centered computing" on "visualization".

4.

$$P(picsimulation \mid \text{do}(collisionlessplasma)) = P\left(picsimulation \mid collisionlessplasma\right)$$

'collisionless plasma' directly affects the keyword 'pic simulation' without any mediators or confounders.

5.

$$P(automation \mid \text{do}(airtrafficcontrol)) = P\left(automation \mid airtrafficcontrol\right)$$

In this relationship there are also no other variables, indicating a direct relationship of "automation" given an intervention in "air traffic control".

5.2.2 Computer and Information Science

5.2.2.1 Vector Autoregression

The corresponding dataset contains 5595 publications which span from 1980 to the present. The total number of unique keywords in this dataset is 8103, out of which 220 pair combinations were selected for analysis of 220 unique keywords.
Using the stationarity testing for variables with a maximum of one time differencing, 129 stationary time series were selected and 8155 equations were generated, the results from six equations are shown below:

1. For variables 'andersson-madigan-perlman interpretation' and 'temporal networks' the number of observations is 40 and the smallest AIC score belongs to lag 5
   For the 'andersson-madigan-perlman interpretation' equation:

$$AMP_t = 0.966667 \times AMP_{t-1} + \epsilon_{AMP}$$

   For the 'temporal networks' equation:

$$\begin{aligned} TN_t = {} & 3.000000 \times AMP_{t-1} - 1.000000 \times TN_{t-1} + 2.000000 \times AMP_{t-2} \\ & + 1.000000 \times AMP_{t-3} \\ & + 1.000000 \times AMP_{t-5} + \epsilon_{TN} \end{aligned}$$

2. For variables 'approximate reasoning' and 'middleware' the number of observations is 40 and the smallest AIC score belongs to lag 5
   For the 'approximate reasoning' equation:

$$\begin{aligned} AR_t = {} & 0.292086 \times MW_{t-1} \\ & - 0.223762 \times AR_{t-2} \\ & + 0.397143 \times MW_{t-4} \\ & + 0.930008 \times MW_{t-5} + \epsilon_{AR} \end{aligned}$$

   For the 'middleware' equation there are no statistically significant components, so we don't have an equation for it.

3. For variables 'approximate reasoning' and 'modelicaml' the number of observations is 40 and the smallest AIC score belongs to lag 5
   For the 'approximate reasoning' equation there are no statistically significant components. For the 'modelicaml' equation:

$$ML_t = 0.449162 \times ML_{t-1} + 0.193978 \times AR_{t-3} + 0.280663 \times AR_{t-5} + \epsilon_{ML}$$

4. For variables 'emergency management' and 'geographical information systems' the number of observations is 40 and the smallest AIC score belongs to lag 5
   For the 'emergency management' equation:

$$\begin{aligned} EM_t =& 0.091945 - 2.065842 \times EM_{t-1} + 2.744456 \times GIS_{t-1} - 1.244918 \times EM_{t-2} + 3.409288 \times GIS_{t-2} \\ & - 0.694338 \times EM_{t-3} + 2.502973 \times GIS_{t-3} - 0.015254 \times EM_{t-4} + 0.478714 \times GIS_{t-4} \\ & - 0.305826 \times EM_{t-5} + 0.230643 \times GIS_{t-5} + \epsilon_{EM} \end{aligned}$$

   For the 'geographical information systems' equation:

$$\begin{aligned} GIS_t =& 0.086027 - 1.854152 \times EM_{t-1} + 1.796192 \times GIS_{t-1} - 0.791616 \times EM_{t-2} + 2.458303 \times GIS_{t-2} \\ & - 0.170515 \times EM_{t-3} + 1.844482 \times GIS_{t-3} + 0.379438 \times EM_{t-4} - 0.320234 \times GIS_{t-4} \\ & - 0.113950 \times EM_{t-5} - 0.533134 \times GIS_{t-5} + \epsilon_{GIS} \end{aligned}$$

5. For variables 'peer-to-peer' and 'security' the number of observations is 42 and the smallest AIC score belongs to lag 3
   For the 'peer-to-peer' equation:

$$\begin{aligned} P2P_t =& 0.114392 \\ & + 0.163826 \times P2P_{t-1} \\ & - 0.030420 \times S_{t-1} \\ & + 0.054124 \times P2P_{t-2} \\ & + 0.190504 \times S_{t-2} \\ & + 0.128975 \times P2P_{t-3} \\ & - 0.191939 \times S_{t-3} \\ & + \epsilon_{P2P} \end{aligned}$$

   For the 'security' equation:

$$\begin{aligned} S_t =& 0.481546 \\ & + 0.160326 \times P2P_{t-1} \\ & + 0.322025 \times S_{t-1} \\ & + 0.034189 \times P2P_{t-2} \\ & + 0.188709 \times S_{t-2} \\ & - 0.991015 \times P2P_{t-3} \\ & + 0.134024 \times S_{t-3} \\ & + \epsilon_S \end{aligned}$$

6. For variables 'computer science' and 'information storage and retrieval systems' the number of observations is 44 and the smallest AIC score belongs to lag 1
   For the 'computer science' equation:

$$\begin{aligned} CS_t =& 0.419914 \\ & + 0.586123 \times CS_{t-1} \\ & - 1.016162 \times ISRS_{t-1} \\ & + \epsilon_{CS} \end{aligned}$$

   For the 'information storage and retrieval systems' equation:

$$\begin{aligned} ISRS_t =& 0.119517 \\ & + 0.235977 \times CS_{t-1} \\ & - 0.394773 \times ISRS_{t-1} \\ & + \epsilon_{ISRS} \end{aligned}$$

The first three equations contain keyword pairs that do not co-occur among the top ranking of the dataset, while the latter three correspond to the valid keyword pairs from the dataset. The first 2 sets show approximately zero residual correlation indicating less linear relation. while the other pairs exhibit correlation values higher than 0.70 specifically:

- The correlation between **approximate reasoning** and **modelicaml** is 0.710612.
- The correlation between **computer science** and **information storage and retrieval systems** is 0.871716.
- The correlation between **peer-to-peer** and **security** is 0.757799.
- The correlation between **emergency management** and **geographical information systems** is 0.922432.

For the highly correlated pair at the time lag 5, figure 5.23 shows the 20 future timesteps. According to this figure the variable 'emergency management' starts from a value of 0.09 with an increasing trend towards a value of 0.3 as its stable state. On the other hand, the keyword 'geographical information systems' follows a slightly different pattern, The series starts at a value of 0.086. There is a slight dip to 0.070 at the second timestep, followed by an increase to 0.095 at the third timestep.
From the third to the fifth timestep, there is another increase, reaching a value of 0.15. The value continues to rise, peaking at 0.184 at the tenth timestep.
After the peak, the value slightly decreases and fluctuates around 0.18 from the 11th to the 16th timestep. Finally, the series ends at a value of 0.17571997, which is higher than the starting value.
As noted, each keyword variable in this dataset follows slightly different trends during the time period of publications.

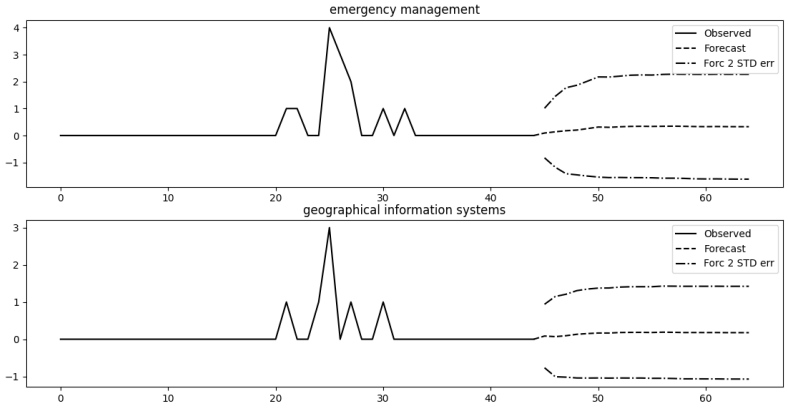


Figure 5.23: 20 Future timesteps for 'emergency management', 'geographical information systems' keywords

5.2.2.2 Granger Causality

Similar to the previous section, for each equation in a VAR system, using the Granger causality test, we can check the hypotheses that each of the keywords in the data can Granger-cause the target variable in that equation. In this section granger causality was performed on the most frequent keywords in this dataset. Table 5.14,5.15 show the results of the Granger causality test on 30 keywords with a time lag of 5.
from the first table we have :

1. **Quality of Service (QoS) and Real-Time Systems**: The p-values for lags 1, 3, 4, and 5 are all above 0.05, suggesting that changes in "real-time system" do not Granger-cause changes in "QoS" at these lags. However, the p-value for lag 2 is below 0.05, indicating that the "real-time system" does Granger-cause "QoS" at lag 2.

2. **Bullying and Moral Disengagement**: The p-values for lags 2, and 3 are above 0.05, suggesting that "moral disengagement" does not Granger-cause "bullying" at these lags. However, the p-values for lags 1, 4, and 5 are below 0.05, indicating that "moral disengagement" does Granger-cause "bullying" at these lags.

3. **Software-Defined Networking (SDN) and Security**: The p-values for all lags are both above 0.05, suggesting that "security" does not Granger-cause "SDN" at these lags. The occurrences in the years corresponding to lags do not seem to have a significant influence on "SDN".

4. **Modeling and Simulation**: The p-value for lag 1, 4, and 5 are above 0.05, suggesting that "simulation" does not Granger-cause "modeling" at these lags. However, the p-values for lags 2 and 3 are below 0.05, indicating that "simulation" does Granger-cause "modeling" at these lags.

5. **(SoC) and System-on-Chip**: The p-values for lags 2 to 5 are below 0.05, indicating that "system-on-chip" does Granger-cause "SoC" at only these lags. This is unexpected as "SoC" and "system-on-chip" are the same concept, just expressed differently.

and from the second table, we can conclude :

1. **Real-time system and QoS**: Only the p-value for lag 3 is less than 0.05, suggesting a granger causation in this time lag. Overall the p-values for lags 1 to 5 range from 0.0227 to 0.0853. This suggests that the keyword "QoS" might have some influence on the keyword "real-time system" in the following years, but the evidence is not very strong as all p-values are greater than 0.05.

2. **Moral disengagement and Bullying**: The p-values for lags 1 to 5 range from 0 to 0.0846. The p-values at lag 1, lag 4, and 5 are 0.0395, 0.0001, and 0 respectively, which are less than 0.05, indicating strong evidence that "bullying" Granger causes "moral disengagement".

3. **Security and SDN**: The p-values for lags 1 to 5 range from 0.2851 to 0.9781. These high p-values suggest that "SDN" does not Granger-cause "security".

4. **Simulation and Modeling**: The p-values for lags 1 to 5 range from 0.0616 to 0.3704. These p-values suggest that "modeling" might not have a significant influence on "simulation".

5. **System-on-chip and SoC**: The p-values for lags 1 to 5 range from 0.0316 to 0.9268. The p-value at lag 5 is 0.0316, which is less than 0.05, indicating that "SoC" might Granger-cause "system-on-chip" at a lag of 5 years.

Table 5.14: Granger Causality test - The p-values are based on the F-test for the Department of Computer and Information Science[Keyword 2 Granger causes Keyword 1]

Keywords 1	Keywords 2	lag 1	lag 2	lag 3	lag 4	lag 5
qos	real-time system	0.3529	0.0178	0.1248	0.093	0.0877
bullying	moral disengagement	0.0112	0.2007	0.6155	0.0001	0.0004
sdn	security	0.4092	0.4255	0.0625	0.0822	0.1314
modeling	simulation	0.238	0.0021	0.0196	0.0636	0.0535
soc	system-on-chip	0.1227	0.0017	0	0	0
computer communication networks	computer science	0.0248	0.1032	0.2397	0.3191	0.0017
hip	security	0.1556	0.0281	0.0359	0.0075	0.0451
privacy	security	0.7034	0.4742	0.112	0.0821	0.0562
declarative programming	program correctness	0.1494	0.3574	0.051	0.0229	0.0275
parallel computing	software composition	0.5756	0.5193	0.6126	0.2023	0.0312
control system	propagation	0.633	0.7763	0.035	0.6528	0.6047
modeling	petri nets	0.8588	0.5894	0.7413	0.9097	0.9298
bullying	bystander	0.1026	0.0561	0.0024	0	0.0003
declarative programming	logic programming	0.8164	0.4913	0.6526	0.8289	0.6712
control system	safety-critical	0.6339	0.7773	0.0351	0.6518	0.6029
design optimization	fault tolerance	0.3353	0.0122	0.0487	0.0885	0.0798
plm	product lifecycle management	0.2308	0.9312	0.2105	0.0144	0.029
java	security	0.0012	0.0021	0.0056	0.0004	0.0028
datavetenskap	festskrift	0.106	0.2844	0.4185	0.6232	0.7837
propagation	safety-critical	0.4163	0.3814	0.5889	0.7072	0.3707
design for testability	testing	0.231	0.0727	0.1822	0.3179	0.7008
formal verification	real-time systems	0.2162	0.4428	0.4522	0.5102	0.781
formal verification	modeling	0.182	0.4617	0.5187	0.6169	0.5411
computer science	festskrift	0.9255	0.2117	0.4056	0.6031	0.3108
bullying	classroom climate	0.2004	0.1036	0.2709	0.0058	0.0004
computer science	datavetenskap	0.9843	0.9756	0.9949	0.997	0.8706
bullying	victimization	0.2167	0.0001	0	0	0.0002
logic programming	program correctness	0.5593	0.8508	0.9098	0.0342	0.0675
multicore processor	parallel computing	0.0082	0.0114	0.0095	0.011	0.1457
fault tolerance	middleware	0.0558	0.045	0.0326	0.0134	0.0055

Table 5.15: Granger Causality test in reverse - The p-values are based on the F-test for the Department of Computer and Information Science [Keyword 2 Granger causes Keyword 1]

Keywords 1	Keywords 2	lag 1	lag 2	lag 3	lag 4	lag 5
real-time system	qos	0.056	0.0853	0.0227	0.0448	0.0793
moral disengagement	bullying	0.0395	0.0846	0.0514	0.0001	0
security	sdn	0.9781	0.7419	0.8075	0.4047	0.2851
simulation	modeling	0.0616	0.2587	0.3704	0.1955	0.2954
system-on-chip	soc	0.7721	0.904	0.9268	0.4516	0.0316
computer science	computer communication networks	0.0044	0.0171	0.0521	0.0785	0.0003
security	hip	0.4008	0.5745	0.4443	0.1947	0.0645
security	privacy	0.1663	0.432	0.2616	0.2837	0.345
program correctness	declarative programming	0.2489	0.5078	0.5126	0.8109	0.8244
software composition	parallel computing	0.0761	0.2392	0.1217	0.1521	0.1143
propagation	control system	0.003	0.0049	0	0	0
petri nets	modeling	0.5157	0.7629	0.8641	0.9204	0.793
bystander	bullying	0.0266	0.1236	0.0004	0	0
logic programming	declarative programming	0.4888	0.7645	0.791	0.4865	0.5489
safety-critical	control system	0.003	0.0049	0	0	0
fault tolerance	design optimization	0.1965	0.1939	0.2436	0.24	0.3899
product lifecycle management	plm	0.2218	0.8468	0.6334	0.9145	0.671
security	java	0.905	0.8806	0.7184	0.5623	0.6906
festskrift	datavetenskap	0.8082	0.9587	0.9903	0.9972	0.9991
safety-critical	propagation	0.8292	0.8236	0.4982	0.4716	0.4538
testing	design for testability	0.8042	0.3096	0.6256	0	0
real-time systems	formal verification	0.8247	0.4031	0.3601	0.0372	0.0074
modeling	formal verification	0.5417	0.4457	0.9143	0.3959	0.355
festskrift	computer science	0.3594	0.0805	0.1928	0.3204	0.1443
classroom climate	bullying	0.3637	0.0036	0.0087	0.0313	0.0222
datavetenskap	computer science	0.3979	0.6883	0.8206	0.904	0.9506
victimization	bullying	0.0099	0.0115	0.0001	0	0
program correctness	logic programming	0.7123	0.9451	0.9956	0.9327	0.8433
parallel computing	multicore processor	0.7923	0.3465	0.4229	0.076	0.1166
middleware	fault tolerance	0.9485	0.8203	0.944	0.9511	0.9818

5.2.2.3 Bayesian Network Analysis

Much similar to 5.2.1.3, In this section two Bayesian networks are applied to the dataset corresponding to the Department of Computer and Information Science. The network was implemented on 77 keywords with 80 co-occurrence relationships in between. According to the 5.16 and 5.17 most of the top keywords have higher probabilities of not being present and there is a consistent pattern that signifies such occurrence among the nodes. By looking at the data it can be observed that among the top keywords, the marginal probability of not occurring ranges between 68% to 87%, and the probability of occurrence would change between 13% to 31%.

Table 5.16: Marginal Probabilities for top keywords in Designed Network

Keywords	**0**	**1**
Scheduling	0.7222	0.2778
Service Design	0.8135	0.1865
Knowledge Representation	0.8637	0.1363
Real-Time Systems	0.7209	0.2791
Synchronization	0.8011	0.1989
Skepu	0.8184	0.1816
Artificial Intelligence	0.8618	0.1382
Concurrency	0.8184	0.1816
TAM	0.8004	0.1996
Debugging	0.8664	0.1336
Prototyping	0.8671	0.1329
Propagation	0.7403	0.2597

Table 5.17: Marginal Probabilities for top keywords in Learned Network

Keywords	**0**	**1**
Scheduling	0.9919	0.0081
Service Design	0.9899	0.0101
Knowledge Representation	0.9937	0.0063
Real-time Systems	0.9892	0.0108
Synchronization	0.9985	0.0015
Skepu	0.9978	0.0022
Artificial intelligence	0.9910	0.0090
Concurrency	0.9978	0.0022
TAM	0.9976	0.0024
Debugging	0.9972	0.0028
Prototyping	0.9983	0.0017
Propagation	0.9988	0.0012

Looking at the designed network 5.24, we can observe several nodes as ancestors of a small number of children for each. Trending keywords such as 'embedded systems', 'modelica', or 'testing' serve as a small hub for the corresponding group of nodes. Overall, the designed network for this dataset resembles small island-like clusters with local connections between the nodes. On the other hand, fig 5.25 shows a more hierarchical structure. On the left section of this network, a cluster including the relevant terms of parallel programming and processing related keywords is formed. This graph has a clearer structure and is superior due to clusters with many internal connections suggesting that the keywords are closely related or often co-occur in the same papers.

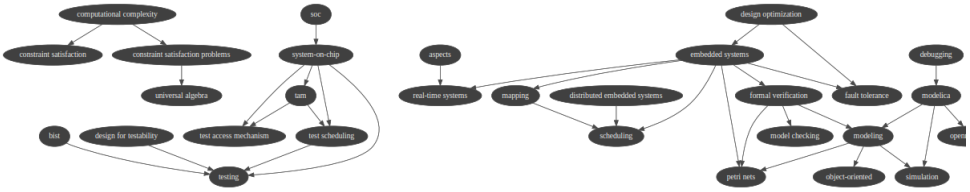


Figure 5.24: Designed structure based on keyword co-occurrence

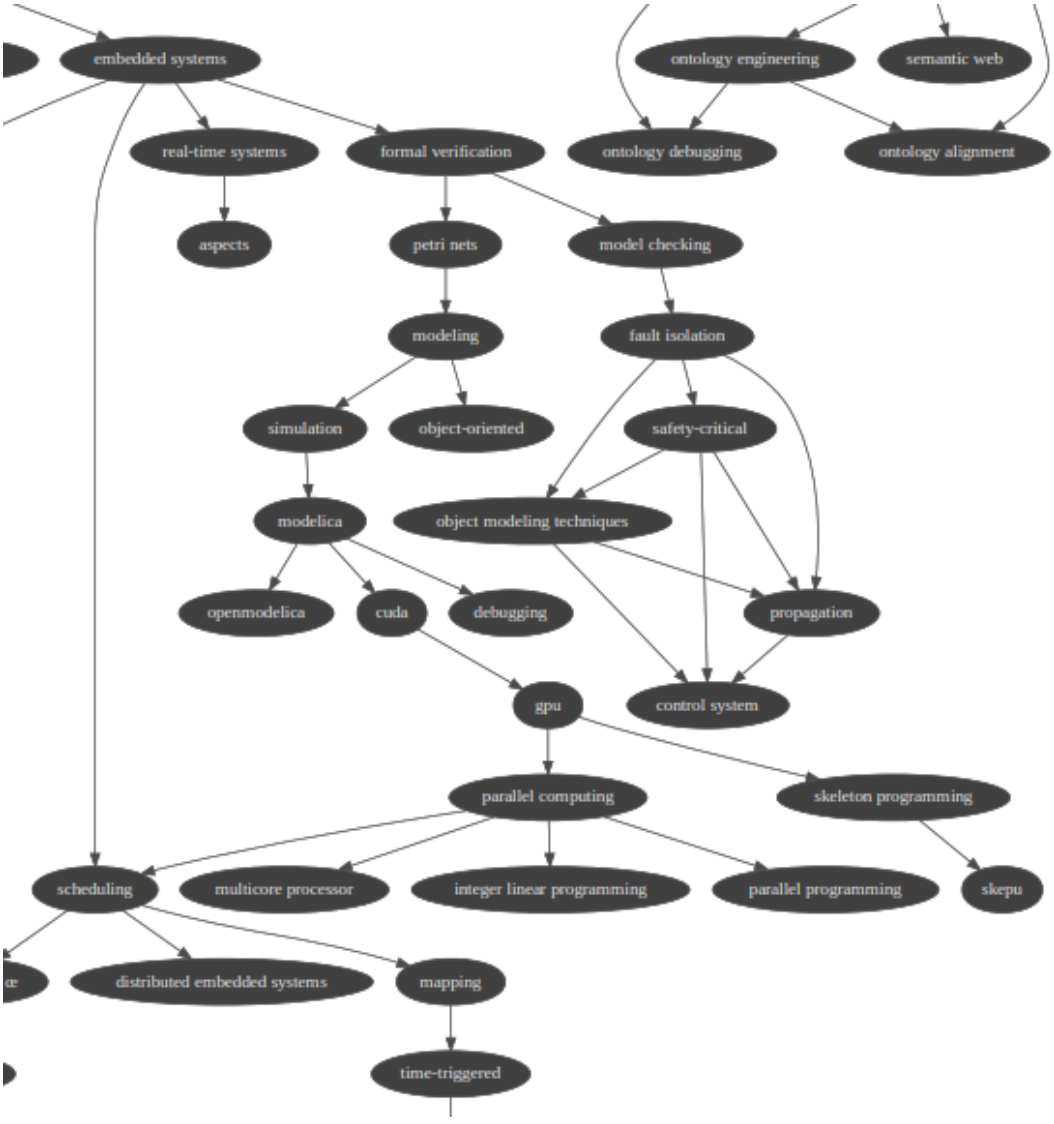


Figure 5.25: Structure learned by MIIC algorithm based on BDeu score

For a more comprehensive visualization refer to 8.2.
Looking at the conditional probabilities for the designed network 5.18, it can be observed that

given the occurrence of evidence, the relevant keyword is less likely to appear. For instance, given that "simulation" is present, "modelica" only has a 0.3517 chance of occurring. The last seven rows are the reverse of the first seven rows. From these results it can be seen that by switching the node and its evidence, the chance of occurring would become random, For example, given "system-on-chip", "testing" has a 0.5010 chance of occurring.
The values observed in the conditional probabilities of the second network 5.19 follow the same pattern as the first network and for many nodes it provides slightly lower probability, indicating a slightly weaker relationship, for example, the probability in the second network (0.3182) is slightly lower than in the first network (0.3364) for Embedded Systems & Real-time Systems pair. Interestingly for nodes 'Bullying & Moral Disengagement' the second network, predicts a significantly higher probability (0.7821) compared to the first network (0.3013), indicating a much stronger relationship in the second network.

Table 5.18: Conditional probabilities of keyword nodes given evidence for designed network

Node Name	**Evidence**	**P(0)**	**P(1)**
system-on-chip	testing	0.7104	0.2896
modelica	simulation	0.6483	0.3517
modeling	simulation	0.5084	0.4916
embedded systems	real-time systems	0.6636	0.3364
test scheduling	testing	0.6209	0.3791
mapping	scheduling	0.6370	0.3630
bullying	moral disengagement	0.6987	0.3013
Testing	System-on-chip	0.4990	0.5010
Simulation	Modelica	0.5041	0.4959
Simulation	Modeling	0.5018	0.4982
Real-time Systems	Embedded Systems	0.5050	0.4950
Testing	Test Scheduling	0.5010	0.4990
Scheduling	Mapping	0.4993	0.5007
Moral Disengagement	Bullying	0.4994	0.5006

Table 5.19: Conditional probabilities of keyword nodes given evidence for learned network

Node Name	**Evidence**	**P(0)**	**P(1)**
system-on-chip	testing	0.6837	0.3163
modelica	simulation	0.6478	0.3522
modeling	simulation	0.6652	0.3348
embedded systems	real-time systems	0.6818	0.3182
test scheduling	testing	0.7789	0.2211
mapping	scheduling	0.6429	0.3571
bullying	moral disengagement	0.2179	0.7821
testing	system-on-chip	0.3261	0.6739
simulation	modelica	0.7245	0.2755
simulation	modeling	0.5674	0.4326
real-time systems	embedded systems	0.7695	0.2305
testing	test scheduling	0.5886	0.4114
scheduling	mapping	0.3364	0.6636
moral disengagement	bullying	0.4020	0.5980

Next, as in the previous dataset, the causal impact in both the designed and the learned networks is computed. The results for the top 5 and their reverse form are as follows for the designed network:

1.

$$P(\text{system-on-chip} \mid \text{do}(\text{testing})) = \sum_{\text{design for testability}} \left\{ P\left(\text{design for testability} \mid \text{testing}\right) \cdot \left(\sum_{\text{testing}'} P\left(\text{system-on-chip} \mid \text{design for testability}, \text{testing}'\right) \cdot P\left(\text{testing}'\right) \right) \right\}$$

In this case, the frontdoor variable is 'design for testability'. This variable is the mediator that is affected by "testing" (the treatment) and in turn affects "system-on-chip" (the outcome).

2.

$$P(\text{modelica} \mid \text{do}(\text{simulation})) = \sum_{\text{aspects}} P(\text{aspects}) \cdot \left(\sum_{\text{simulation}'} P(\text{modelica} \mid \text{simulation}') \cdot P(\text{simulation}') \right)$$

"simulation" keyword influences the presence of "modelica" keyword indirectly through 'aspects'.

3.

$$P(\text{modeling} \mid \text{do}(\text{simulation})) = \sum_{\text{aspects}} P(\text{aspects}) \cdot \left(\sum_{\text{simulation}'} P(\text{modeling} \mid \text{simulation}') \cdot P(\text{simulation}') \right)$$

Similar to the second expression, "simulation" keyword also influences "modeling" indirectly through 'aspects'.

4.

$$P(\text{embedded systems} \mid \text{do}(\text{real-time systems})) = \sum_{\text{debugging}} P(\text{debugging}) \cdot \left(\sum_{\text{real-time systems}'} P(\text{embedded systems} \mid \text{real-time systems}') \cdot P(\text{real-time systems}') \right)$$

The frontdoor variable 'debugging' keyword is a mediator that is affected by "real-time systems" (the treatment) and in turn affects "embedded systems" (the outcome).

5.

$$P(\text{test scheduling} \mid \text{do}(\text{testing})) = \sum_{\text{design for testability}} \left[P(\text{design for testability} \mid \text{testing}) \cdot \left(\sum_{\text{testing}'} P(\text{test scheduling} \mid \text{design for testability}, \text{testing}') \cdot P(\text{testing}') \right) \right]$$

Lastly the frontdoor variable is 'design for testability' which acts as a mediator for "testing" (the treatment) and "test scheduling" (the outcome).

Given the reverse case for designed network:

1.
$$P(\text{testing} \mid \text{do}(\text{system-on-chip})) = \sum_{\text{bist,design for testability,tam,test scheduling}} \Big[P(\text{bist}) \cdot P(\text{test scheduling} \mid \text{system-on-chip}, \text{tam}) \cdot P(\text{design for testability}) \cdot P(\text{tam} \mid \text{system-on-chip}) \cdot P(\text{testing} \mid \text{bist}, \text{design for testability}, \text{system-on-chip}, \text{test scheduling}) \Big]$$

To compute the causal effect of "system-on-chip" on "testing", using the do-calculus, the formula needs to take into account 'bist', 'design for testability', 'tam', and 'test scheduling variables.

2.
$$P(\text{simulation} \mid \text{do}(\text{modelica})) = \sum_{\text{design optimization,embedded systems,formal verification,modeling}} P(\text{modeling} \mid \text{formal verification}, \text{modelica}) \cdot P(\text{embedded systems} \mid \text{design optimization}) \cdot P(\text{simulation} \mid \text{modelica}, \text{modeling}) \cdot P(\text{formal verification} \mid \text{embedded systems}) \cdot P(\text{design optimization})$$

Similar to the first equation the model the presence of the keyword "simulation"(adjusted causal effect of "modelica" on "simulation") is dependent on 4 other keywords 'design optimization', 'embedded systems', 'formal verification', and 'modeling'.

3.
$$P(\text{simulation} \mid \text{do}(\text{modeling})) = \sum_{\text{modelica}} \left[P\left(\text{simulation} \mid \text{modelica}, \text{modeling}\right) \cdot P\left(\text{modelica}\right) \right]$$

Through this equation "modelica" is the confounder for the causal relationship of 'modeling' on 'simulation'.

4.
$$P(\text{real-time systems} \mid \text{do}(\text{embedded systems})) = \sum_{\text{aspects}} P\left(\text{real-time systems} \mid \text{aspects}, \text{embedded systems}\right) \cdot P\left(\text{aspects}\right)$$

To measure the probability of "real-time systems" given an intervention on "embedded systems", the value of the 'aspects' keyword is needed to be considered.

5.
$$P(\textit{testing} \mid \text{do}(\textit{test scheduling})) = \sum_{\textit{system-on-chip}} \Big(P(\textit{testing} \mid \textit{system-on-chip}, \textit{test scheduling}) \cdot P(\textit{system-on-chip}) \Big)$$

In this equation, 'system-on-chip' is a backdoor variable and therefore a confounder keyword that affects both "test scheduling" (the treatment) and "testing" (the outcome).

Through the learned network, the causal effects can be computed as:

1.

$$P(\textit{system-on-chip} \mid \text{do}(\textit{testing})) = \sum_{\textit{aspects}} \Big(P(\textit{aspects} \mid \textit{testing}) \cdot \Big(\sum_{\textit{testing}'} P(\textit{system-on-chip} \mid \textit{testing}') \cdot P(\textit{testing}') \Big) \Big)$$

The first equation signifies the frontdoor effect of 'aspects' in the causal relationship of "testing" on "system-on-chip".

2.

$$P(\textit{modelica} \mid \text{do}(\textit{simulation})) = \sum_{\textit{testscheduling}} P\left(\textit{modelica} \mid \textit{simulation}\right) \cdot P\left(\textit{testscheduling}\right)$$

'test scheduling' is the backdoor for the relationship of "simulation" (the treatment) and "modelica" (the outcome).

3.

$$P(\textit{modeling} \mid \text{do}(\textit{simulation})) = \sum_{\text{multicore processor}} \Big(P(\textit{multicoreprocessor} \mid \textit{simulation}) \cdot \Big(\sum_{\textit{simulation}'} \Big(P(\textit{modeling} \mid \textit{simulation}') \cdot P(\textit{simulation}') \Big)$$

In this equation 'multicore processor' acts as the frontdoor variable for the causal effect of "simulation" on "modeling".

4.

$$P(\textit{embedded systems} \mid \text{do}(\textit{real-time systems})) = \sum_{\textit{aspects}} \Big(P(\textit{aspects} \mid \textit{real-time systems}) \cdot \Big(\sum_{\textit{real-time systems}'} \big(P(\textit{embedded systems} \mid \textit{real-time systems}') \cdot P(\textit{real-time systems}') \big) \Big) \Big)$$

In the relationship between "real-time systems" (the treatment) and "embedded systems" (the outcome), 'aspects' keyword act as the frontdoor variable.

5.

$$P(\textit{test scheduling} \mid \text{do}(\textit{testing})) = \sum_{\textit{aspects}} \Big(P(\textit{aspects} \mid \textit{testing}) \cdot \Big(\sum_{\textit{testing}'} \Big(P(\textit{test scheduling} \mid \textit{testing}') \cdot P(\textit{testing}') \Big) \Big) \Big)$$

similar to the third equation, 'aspects' keyword acts as the frontdoor variable for the causal effect of "testing" on "test scheduling".

As for the reverse case:

1.

$$P(\textit{testing} \mid \text{do}(\textit{system-on-chip})) = \sum_{\textit{test scheduling}} \Big(P(\textit{test scheduling} \mid \textit{system-on-chip}) \cdot P(\textit{testing} \mid \textit{system-on-chip}, \textit{test scheduling}) \Big)$$

To compute the probability of "testing" given intervention on "system-on-chip", the summation of all possible states of "test scheduling" is needed.

2.

$$P(simulation \mid \text{do}(modelica)) = \sum_{\text{multicore processor}} \Big(P(multicoreprocessor \mid modelica) \cdot \Big(\sum_{modelica'} \Big(P(simulation \mid modelica') \cdot P(modelica') \Big) \Big) \Big)$$

In this equation, 'multicore processor' is a front-door variable and mediator in the causal pathway from modelica to simulation.

3.

$$P(simulation \mid \text{do}(modeling)) = \sum_{test\ scheduling} \Big(P(simulation \mid modeling) \cdot P(test\ scheduling) \Big)$$

Through the backdoor criterion, 'test scheduling' is estimated to be the confounder for both 'modeling' (treatment) and simulation (outcome).

4.

$$P(real-time\ systems \mid \text{do}(embedded\ systems)) = \sum_{testing} \Big(P(real-time\ systems \mid embedded\ systems) \cdot P(testing) \Big)$$

From this equation 'testing' is the backdoor variable for the causal effect of 'embedded systems' on 'real-time systems'.

5.

$$P(testing \mid \text{do}(test\ scheduling)) = \sum_{\text{system-on-chip}} P(testing \mid \text{system-on-chip, test scheduling}) \cdot P(\text{system-on-chip})$$

In this case (Backdoor criterion) it is also revealed that 'system-on-chip' is a backdoor variable and a confounder in the causal pathway from 'test scheduling' to 'testing' keywords.

6 Discussion

In this chapter, a summary of the challenges faced during the implementation of methods, the results associated with the methodology, and the future vision for this project are discussed. 6.1 briefly describes the methods and the resulting conclusions, similar methods in articles, 6.2 provides a description of limitations of this approach during the actual implementation, 6.3 mentions some ideas about the future directions of research and potential approaches that can help with capturing more information from scientometric datasets.

6.1 Method

The keywords serve as a concise overview of the principal topics and findings in the document. They provide a quick insight into the main research themes and content, allowing for easy identification of the document's core subjects and discoveries. This thesis aimed to investigate the author-defined keyword interactions in scientific publications. Particularly, frequent keywords in the mentioned publications were the focus of this study.
For all the methods used, the data was processed beforehand. This improves the overall quality of the dataset and allows for a more comprehensive keyword analysis for both trend observation 4.3 and influence analysis. As a first attempt, the linear regression method on the top keywords. Regression models are generally used to predict trending behaviors and theoretically, they can be helpful when studying the evolution of a given series[4], in our case however, a generally upward trend for the keywords under study with very few exceptions. The primary cause for such results is that from the point of emergence for keywords, they usually represent a gradual increase over the initial years of occurrence where some of them might experience an era of decline afterward, moreover, the non-linearity in the behavior of keywords compelled us to consider alternative techniques. experimenting with the data against univariate distributions through a statistical test of Kolmogorov-Smirnov(K-S test). In the field of scientometrics, while this method is not often applied, researchers employ other statistical methods such as descriptive statistics, regression-based methods [124], and a combination of methods [111]; unfortunately, most of such works solely focus on the analysis of citation metrics or designing a form of citation analysis. In this work, the K-S test determines keyword frequency distribution by experimenting with 10 theoretical univariate distributions. Through statistical significance, this test was able to detect likely candidates for the distribution that covers the keywords within the dataset, and among the tested keywords

the t distribution was the most prevalent one.
The other method to observe the trend in the dataset was to prepare each dataset so that it only contains keywords within the past five years, furthermore, all previously existing keywords were discarded and the final prepared version contained only the appropriate keywords for inference. Scanning individual datasets resulted in different layouts of topics based on the probabilities of words contained within. Specifically, the number of topics is the deciding factor that can generate coherent topics with clusters of similar keywords with respect to the articles 5.19 or simply generate closely fitted topics with little to no valuable insights into the connection structure of these keywords. Through these efforts, the algorithm was able to identify specific themes that reflected the prevailing research trends across the three target datasets. Similar works have been published for the analysis of topics across fields, studies such as [87](for online reviews in the food industry), [42](articles in Journal of Applied Intelligence for over 30 years), and [86](1000 interdisciplinary scientific papers from 'PLOS ONE' journal), however, the majority of these works focus on a selected number of topics based on only the meaning of topics and the fit on the corresponding datasets. In this thesis, three metrics were used to assess the quality of topics and choose the best descriptive ones that are both **appropriately different** and **meaningful**.
This thesis also introduced a frequency counter algorithm designed to investigate the co-occurrence of keywords in a dataset using the concept of n-grams. The outcomes generated by this algorithm served as input for the Vector Autoregression and Causality inference methods.
Overall in this study, three methods were utilized:

1. **Vector Autoregression (VAR):** This method was used to understand the interdependencies across multiple time series variables. Each variable in the system is modeled as a linear function of past lags of itself and the past lags of other variables. This helped in capturing the temporal dynamics in the keyword frequencies.

2. **Granger Causality:** This statistical concept was used to find out if one time series is useful in forecasting another. In the context of this research, it was used to determine if the frequency of one keyword can predict the frequency of another keyword in the future.

3. **Bayesian Network Analysis:** This probabilistic graphical model was used to represent a set of variables and their conditional dependencies via a directed acyclic graph. It was used to understand the probabilistic relationships among the keywords.

In regards to multivariate time series, VAR was used successfully involving variables connected in the same field, [96](model of relationship between renewable energy investment and oil prices), [8](exploring the relationship between climate change and agricultural production in the Mediterranean region), [51](explaining the relationship between download and citation counts through grange causality), and [24](Relationship between China's air passenger traffic to three individual factors) are examples of these studies. Through VAR the associated time series of keywords with frequent co-occurrence were analyzed, Thereafter through Granger causality the predictive relationship between keywords within the publication datasets was established. Several of these relationships were examined(5.8,5.9,5.14,5.15), with a confidence level of 95% and the potential time lag of 5 years that the effect of one keyword on another could manifest up to 5 years later were described. Based on these observations, it could be seen that synonym keywords rarely appear together(No G-causal relationship) and some of the apparently connected keywords, do not have a predictive relationship(e.g. privacy and security).
The last methodology applied was Bayesian Network analysis for computing the causal effect of the datasets through parameter learning for the designed network(5.21,5.24) and learning the complete structures(5.22,5.25). These were shown to perform well in literature([64, 82]),

and in this study, the causal relationship including the backdoor and frontdoors of the various keywords was observed, and in some cases, seemingly connected keywords were to show unexpected keywords as mediators or confounders within such relationships(e.g. 'machine learning' being a con-founder of "visual analytics" (the treatment) and "visualization" (the outcome)).

6.2 Limitations

While each of the methods referenced in 6.1 shows promise, still analyzing scientific literature is a challenging task. Scientometrics often involves interdisciplinary approaches [38], combining techniques from fields such as information science and social sciences. Therefore integrating the perspective can be challenging. One problem faced during this keyword analysis study was the diversity of keywords used within the scientific publications. which would result in the inefficiency of the Bayesian network construction for full datasets. Additionally, the challenges faced during the practical phase of this thesis and ultimately some of the limitations are as follows:

1. **Data Preprocessing:** Cleaning the data and ensuring the quality was a significant challenge. The dataset contained missing values and inconsistencies that needed to be addressed before analysis.

2. **Determining the Optimal Lag:** For VAR and Granger Causality, determining the optimal number of lags was a challenge. Various criteria like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used, but choosing the best one required careful consideration. Because of the unequal amount of observations for these time series and consideration of the influence of these keywords, the time lag of 5 years for all the keywords was chosen, but a better choice for such an approach is to use more flexible methods for this task([43, 6]).

3. **Complexity of Bayesian Networks:** The construction and interpretation of Bayesian Networks was complex due to the large number of keywords. Simplifying the network without losing important information was a difficult task.

4. **Computational Resources:** The methods used were computationally intensive, especially with a large number of keywords. This required efficient coding practices and sometimes made the analysis time-consuming.

5. **Keyword Variations:** Dealing with variations and synonyms of keywords is a limitation for the algorithms used here. For instance, different authors might use different terms for the same concept, which needs to be identified and treated as the same keyword, in the case of LDA analysis this can only be remedied if the synonyms appear frequently together which is rarely the case in scientific articles. Similarly for VAR and Bayesian Networks, the representatives for the same concepts are treated as different keywords.

6. **Temporal Analysis:** The time series analysis required stationarity of data, which means the properties of the series do not depend on the time at which the series is observed. Transformations such as differencing and logarithms were often used to stabilize the variance, yet careful considerations need to be met in regard to keywords and their distribution. Another challenge was the lack of presence for some keywords in the multiple years of time series which would lead to the production of negative values during 'differencing', ultimately only one differencing operation was performed for each non-stationary keyword in datasets, and the keywords which could not satisfy the condition afterward, were not considered in the next comparison of VAR and Granger tests

7. **Interpretation of Results:** The interpretation of results from VAR and Granger Causality tests required a deep understanding of the subject matter. The statistical significance might not always translate to practical significance.

6.3 Future work

During the process of this research, we faced challenges in the theoretical and practical aspects of the approaches discussed in the thesis. While dealing with these challenges was time-consuming, some of them proved to be very inspiring, therefore In this section, some potential directions for future studies are discussed:

1. **Diversifying the input data:** In the case of this thesis, the analysis was performed on the large dataset corresponding to the Linköping University publication repository and two sub-datasets. While this analysis revealed some patterns, there were no significant results that could indicate patterns in the relationship between the two sub-datasets and the large dataset. Hence one potential direction for future studies would be to include more departments for the aim of such scientific analysis.
   Additionally applying these methodologies to the different collections of papers could yield different results and change some of the perspectives of the approaches discussed.

2. **Bayesian networks new designs:** In this thesis, one of the methodologies used for causal inference is the Bayesian network. Specifically, the designed network in this research involves the causal effects between the co-occurred keywords in the corresponding dataset, however, this network does not consider the temporal aspects of time series corresponding to keywords into account. One direction for future studies would be to design a dynamic network that captures the causal aspects of such relationships across time and effectively observes the keyword evolution through time. This direction requires some modifications to the frequency counting algorithm described in 4.2.

3. **Investigating the dimensionality reduction methods:** While visualizing the results from the LDA technique(5.5), to draw the intertopic distance map, the Principal component analysis (PCA) method was initially used. Unfortunately, the results from the original application were unsatisfactory due to many overlaps between topics with no identical words as a replacement the t-Distributed Stochastic Neighbor Embedding (a non-linear dimensionality reduction) was used.
   As a future research direction, to find the major cause for this difference in performance, the results from these methods and the culprit for non-linear aspects of keywords need to be investigated in depth.

4. **Investigating the categorical property of keywords:** The variables in this thesis are inherently categorical data. The keywords under study are essentially groups of information categorized or classified based on their respective labels or their associated papers. One future direction for this study is to make use of global optimization [112] to compute the levels for values for qualitative data, in order to incorporate the mixed data types into the causal network of keyword analysis.

6.4 Ethical considerations

In the context of research on scientometrics, focusing on the trend, emergence, and influence of keywords on each other using quantitative methods, several societal and ethical considerations should be taken into account.

6.4.1 Ethical Data Collection and Analysis

In research of scientometrics the collection and analysis of bibliometric data, including keywords, should adhere to ethical standards. One ethical consideration in this area is the possibility of bias in data sampling [117], which can potentially hinder research goals and waste resources.
There are also other ethical standards that researchers should consider while using bibliometrics for research evaluations, in [34] a framework for this aspect, including the values and principles for researchers and addressing the absence of community-wide consensus on ethical standards.
In this study, we ensure the accuracy, completeness, and absence of bias in the bibliometric data collected for this study. The methods used in this study are valid and highly proficient in addressing the research questions at hand. Finally, we follow the principle of distributing credit fairly.

7 Conclusion

In this case, four questions regarding the trend, emergence, and influence of keywords were considered. The first two questions were answered through trend analysis by linear regression and clustering through topic modeling, while for the latter two questions, three statistical methods, namely Vector Autoregression, Granger causality testing, and Bayesian Network analysis, were employed.

We utilized the Granger causality method to conduct predictive analysis on the most frequently occurring keywords. Additionally, Bayesian networks were specifically employed to model the presence of keywords within scientific papers. Notably, in order to assess the causal impact of keywords, we employed two distinct networks: one constructed based on expert knowledge and another developed through a hybrid structure learning algorithm. Given the variations in results, both networks were subjected to rigorous testing in each analysis to observe the causal experiments.
All three techniques employed for influence analysis effectively addressed our research questions. Furthermore, within the field of scientometrics, we believe that these approaches hold potential for application to other collections of papers, providing valuable insights into related research domains.

Q1: *What are the most frequently used keywords in scientific literature over the past decades, and how has their usage changed over time?*
The short answer to this question is yes, however, the change in values is inconsistent across keywords due to their different nature and field of study.
To answer this question we refer to the 5.1 table, containing the keywords with the most occurrences in the large dataset. Based on examples of 5.1 and looking at the rest of the frequent keywords, it can be seen that linear regression shows only an upward trend. To demonstrate the trend of change in detail, we can look at the 7.1.

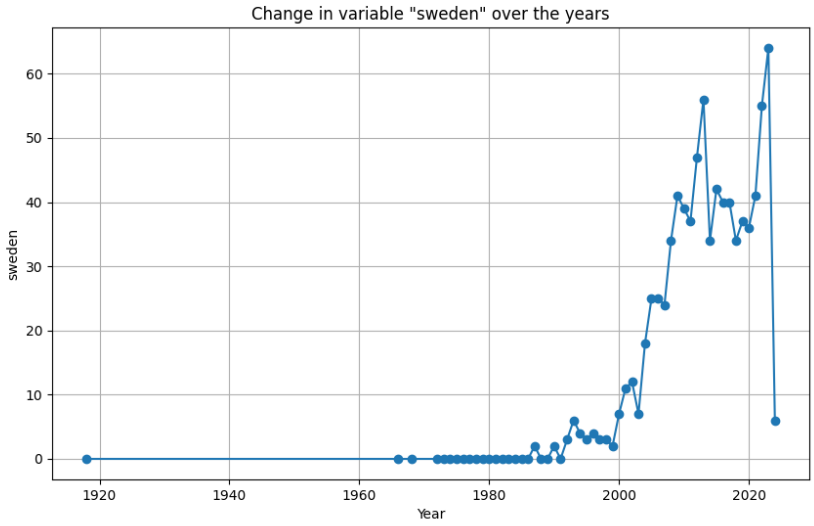


Figure 7.1: Change in the value of Sweden keyword over time

1. **Stable Period (1918 - 1986)**: The value of 'sweden' remained at 0 for a long period from 1918 to 1986. This suggests that 'sweden', had no presence during this period.

2. **Initial Increase (1987 - 1999)**: The first non-zero value appeared in 1987, and the values fluctuated between 0 and 6 until 1999. This period marks the beginning of some activity or influence of 'sweden'. However, the influence seems to be inconsistent during this period as the values are not steadily increasing or decreasing.

3. **Significant Growth (2000 - 2023)**: The period from 2000 to 2023 shows a significant increase in the values, indicating a growing trend of 'sweden'. The value reached a peak of 64 in 2023. This could suggest that 'sweden' has become increasingly important in areas of study over these years.

4. **Sudden Drop (2024)**: The sharp drop to 6 in 2024 is interesting. Since we are currently in 2024, it's possible that this year's data is incomplete, which could explain the sudden drop.

likewise for 'gender' keyword(7.2):

1. **1918-1992: Stability at Zero**
   From 1918 to 1992, the value of the variable 'gender' remains consistently at 0. This long period of stability suggests that gender had no presence during this period.

2. **1993-1995: Initial Increase**
   In 1993, there was a noticeable change, with the value rising to 1 and remaining steady until 1995. This records the first instance of change.

3. **1996-2003: Fluctuations and Gradual Increase**
   The value returns to 0 in 1996, but then fluctuates and begins to increase from 1997 onward. By 2003, the value has risen to 12. This period shows an overall upward trend with some variability, suggesting more impactful events influencing the keyword.

4. **2004-2007: Steady Growth**
   From 2004 to 2007, the value continues to rise steadily, reaching 23 by 2007. This indicates a period of consistent growth in the keyword 'gender'.

5. **2008: Sharp Spike**
   In 2008, there was a dramatic spike to 75, which is a significant outlier compared to previous values. This suggests a major event affecting the keyword.

6. **2009-2024: Volatility and Decline**
   After the peak in 2008, the values fluctuated significantly. There is a general decline from 2009 (60) to 2022 (15), with some intermittent rises and falls. The value drops sharply to 2 in 2024(according to the last update of the dataset).

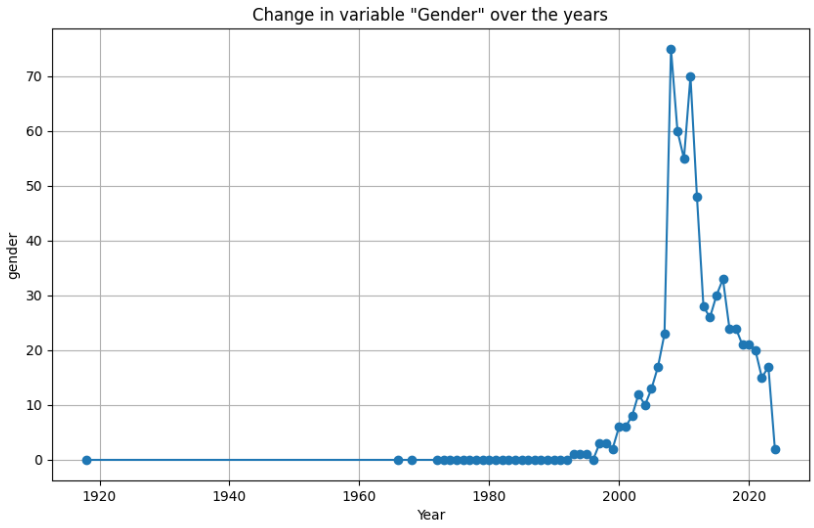


Figure 7.2: Change in the value of Gender keyword over time

As for the sverige keyword(7.3):

1. **Early Period: 1918 - 1979**

   - **1918 - 1979**: The value remains consistently at 0 indicating a prolonged period of no presence.

2. **Initial Increase: 1980 - 1987**

   - **1980**: The value changes to 1, marking the first recorded instance of the keyword.
   - **1981 - 1986**: The value returns to 0, suggesting that the occurrence in 1980 was an anomaly.
   - **1987**: An increase to 2, which continues through 1989. This period shows more frequent occurrences.

3. **Fluctuation and Gradual Rise: 1990 - 1999**

   - **1990**: The value drops back to 0, indicating fluctuations.
   - **1991 - 1999**: There are irregular but generally increasing values, peaking at 7 in 1993 and 2000. This period shows a trend towards more frequent occurrences but with significant variability.

4. **Steady Increase: 2000 - 2005**

   - **2000**: A noticeable jump to 7.

- **2001 - 2005**: Values continue to rise, reaching a peak of 25 in 2004 and maintaining it in 2005 showing a clear upward trend.

5. **Variability and High Values: 2006 - 2024**

   - **2006 - 2024**: This period is distinguished by high variability but generally higher values compared to earlier periods.
     - **2006**: A drop to 9.
     - **2007 - 2014**: Values fluctuate but generally increase, peaking at 46 in 2015.
     - **2015**: Marks the highest value so far at 46.
     - **2016 - 2019**: Values fluctuate but remain high, with a notable drop to 30 in 2019.
     - **2020**: Another peak at 56, the highest in the dataset.
     - **2021 - 2024**: Significant drop-off after 2020, with values decreasing to 3 in 2024.

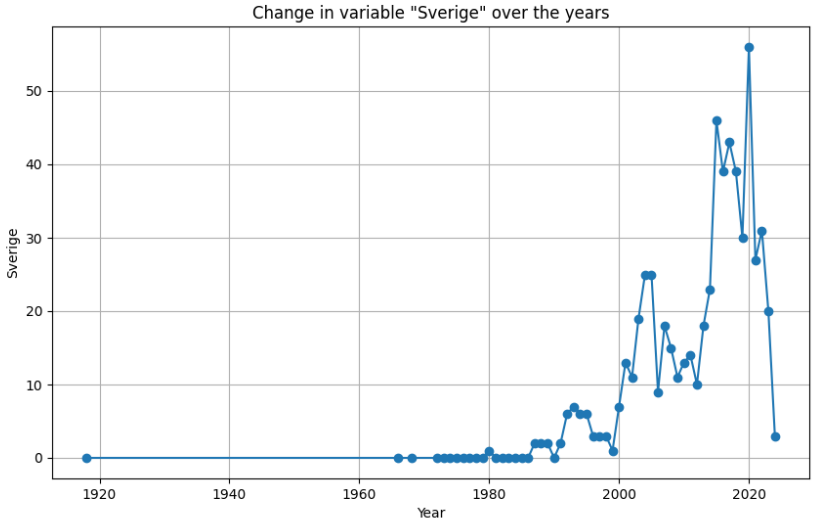


Figure 7.3: Change in the value of Sverige keyword over time

Q2: *Can we identify the themes that have emerged in the last five years from keywords, and what keywords are associated with these themes?*
Yes, we can identify the emerging themes with the methodology. To answer this question the LDA algorithm was applied, this algorithm can categorize different keywords across articles together, effectively clustering for the keywords. The more crucial aspect of this category is the possibility of overlap between clusters, therefore the identified topics of research can contain multiple keywords belonging to articles that best describe them.
This approach resulted in significantly different results, hence the LDA technique was performed at different levels of dataset. For more accuracy in conclusion, three metrics were used to support the decision-making process in addition to the semantic meaning of the generated clusters(e.g. 5.7 and 5.10 vs 5.8 5.11). In the large dataset containing all papers of Linköping University, three prevalent topics observed include advanced materials and technologies, finance, and medical and healthcare. As a result of Computer and Information Science publications, different themes were selected in this section, 'machine learning' being one of the prevalent ones.
In the case of this methodology, computing the topic differences according to the Hellinger

Distance and Jaccard Distance resulted in interesting patterns. Aside from the inherent differences between the datasets under study, it is always observed that as the number of topics grows, the distance between them and therefore the similarity of the topics would increase according to the Jaccard metrics(e.g. 5.11,5.14,5.15). This is because the sets of keywords describing the topics contain more identical words as we increase the number of topics, affecting the measurement. On the other hand, the Hellinger Distance considers the probability distributions of the topics within the corresponding datasets (documents constructed when the datasets were considered as corpus), therefore there is a disparity between the results of each dataset according to basic statistics and even outliers(e.g. 5.12 and 5.16).

Q3: *How does a surge in the usage of a particular keyword influence the usage of other related keywords in scientific publications?*

To answer this question we used Vector autoregression(VAR). Through this methodology, the variables corresponding to keywords can form special mathematical expressions according to the actual co-occurrence in the datasets(4.4). For the purpose of building the most efficient expressions, various time lags for each keyword combination were chosen through Akaike's Information Criterion. An example of the resulting figure and analysis for two keyword-variable of 'shading' and 'shadowing' can be observed in 3.

As an example, we can describe the relationship between two keywords of 'artificial intelligence' and 'automated planning' from the computer and information science dataset by looking at the figure7.4. The overall shape of the prediction for the two variables can be said to follow a similar pattern, but this pattern shows fluctuations when looking into the details of the two keyword-variables. The values indicate the predicted change in each variable at each step. For example, at step 45, the first variable is expected to decrease by approximately 14.36 units, while the second variable is expected to increase by approximately 2.32 units. Similarly, at the 46th observation, the first variable is expected to increase by approximately 3.55 units, while the second variable is expected to increase by approximately 5.33 units. In some steps, there is an increase in one variable while there is a decrease in the other. In the 51st observation variable, artificial intelligence is forecasted to increase significantly to 16.93, while Variable 2 is forecasted to decrease to -6.65. By looking through this process it can be seen that only some steps have the same motion namely among the future steps: 4th, 6th, 8th, 11th, 13th, 15th, and 19th. These instances of the same motion can indicate periods where the variables are reacting similarly to underlying factors, possibly due to their interdependence. When both variables exhibit the same motion, it could indicate that they are being influenced by a common external factor or that there is a strong internal linkage between them causing simultaneous movements.

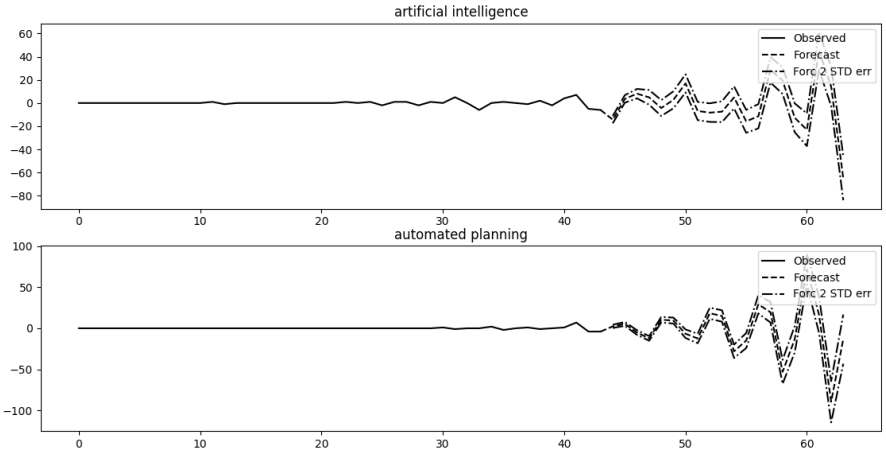


Figure 7.4: artificial intelligence and automated planning forecast for 20 steps into future

Q4: *Is there a general pattern for causal relationships between keywords?*
Based on the observations from the top keywords, it is evident that there is **no** obvious general pattern for causal relationships among keywords. To analyze the causal effects of the keywords in the datasets, granger causality for the temporal aspects and the Bayesian network for the causal impact of keywords with respect to each other.
Through Bayesian networks, we can test various causal effects given the networks(designed and learned). The causal effect under study is primarily to check whether one keyword can predict the occurrence of the other or not. While investigating the causal impact of top keywords (top 30) in the designed and learned networks it was noted that when considering the reverse form of keyword relationship, in the learned Bayesian network backdoor criterion is satisfied more often than in the designed network, in other words, there is a higher chance to discover a confounder keyword through the reverse form of causal impacts. For more comprehensive calculation please refer to 5.2.1.3.

On the other hand, the Granger causality test identifies pairs of keywords with the capability to predict the frequency of the other keyword time series using historical data. As an example to compare the results of these methods consider two keyword variables, according to the tables 5.8 and 5.9, 'visualization' granger-causes 'scientific visualization' on lag 1 to 4, and 'scientific visualization' granger-causes 'visualization' on lag 1 to 3. Computing the causal impact through the designed network:

- $$P(\text{scientific visualization} \mid \text{do}(\text{visualization})) =$$
$$\sum_{\text{data visualization}} \left\{ P\left(\text{data visualization} \mid \text{visualization}\right) \cdot \left(\sum_{\text{visualization'}} P\left(\text{scientific visualization} \mid \text{data visualization}, \text{visualization'}\right) \cdot P\left(\text{visualization'}\right) \right) \right\}$$
which identifies 'data visualization' as a frontdoor for the causal impact of 'visualization' on 'scientific visualization'

- $$P(\text{visualization} \mid \text{do}(\text{scientific visualization})) = \sum_{\text{human-centered computing}} P\left(\text{visualization} \mid \text{human-centered computing}, \text{scientific visualization}\right) \cdot P\left(\text{human-centered computing}\right)$$

  indicating 'human-centered computing' as a backdoor for 'scientific visualization' on 'visualization'.

and for the learned network:

- $$P(\text{scientific visualization} \mid \text{do}(\text{visualization})) = \sum_{\text{human-centered computing}} \left[P\left(\text{human-centered computing} \mid \text{visualization}\right) \cdot \left(\sum_{\text{visualization}'} P\left(\text{scientific visualization} \mid \text{human-centered computing}\right) \cdot P\left(\text{visualization}'\right) \right) \right]$$

  which identifies 'human-centered computing' as a frontdoor for the causal impact of 'visualization' on 'scientific visualization'

- $$P(\text{visualization} \mid \text{do}(\text{scientific visualization})) = \sum_{\text{human-centered computing}} P(\text{visualization} \mid \text{human-centered computing}) \cdot P(\text{human-centered computing})$$

  indicating 'human-centered computing' as a backdoor for 'scientific visualization' on 'visualization'.

The results are consistent for these keywords(Note that conditionals are slightly different). However, this pattern does not always hold, as another example 'visualization' and 'evolution' can not granger-causer each other on any time lags.
If we look at the results from the two networks we have following results:
Designed network:

- $$P(\text{evolution} \mid \text{do}(\text{visualization})) = \sum_{\text{data visualization}} P(\text{data visualization} \mid \text{visualization}) \cdot \left(\sum_{\text{visualization}'} P(\text{evolution} \mid \text{data visualization}, \text{visualization}') \cdot P(\text{visualization}') \right)$$

  The presence of 'evolution' is caused by 'visualization' with frontdoor mediator of 'data visualization'

- The equation for the causal impact of 'evolution' on 'visualization' involves the "data visualization", "deep learning", "digital design", "digital pathology", "evaluation", "haptics", "human-centered computing", "information visualization", "interaction", "interaction design", "life science education", "machine learning", "protein interactions",

"protein-ligand docking", "science education", "science education research", "scientific visualization", "sonification", "text visualization", "virtual reality", and "visual analytics".

we can safely say there is no causal relationship in reverse.
Learned network:

- $$P(evolution \mid \mathrm{do}(visualization)) = P\left(evolution \mid visualization\right)$$

  There is a direct causal effect from 'visualization' on 'evolution' without any frontdoor or backdoor variables.

- $$P(\text{visualization} \mid \mathrm{do}(\text{evolution})) = \sum_{\text{threshold concepts}} P\left(\text{threshold concepts} \mid \text{evolution}\right) \cdot \left(\sum_{\text{evolution}'} P\left(\text{visualization} \mid \text{evolution}'\right) \cdot P\left(\text{evolution}'\right) \right)$$

  'threshold concepts' is a frontdoor variable on the reverse side.

Therefore according to these results the presence of 'evolution' can be caused by 'visualization' in both networks and the presence of 'visualization' can be caused by 'evolution' only in the learned network, but their frequency cannot be predicted based on one another.

8 Appendix

In this chapter, the supplementary results for two sections of the Results chapter are given. The first section provides some results from the statistical test for univariate distributions corresponding to the top keywords(8.1) and in the second(8.2) the full architecture of the Bayesian network for the designed network and learned networks corresponding to the **MIT** and **IDA** datasets are shown.

8.1 Statistical tests for detecting sample distributions

For a better demonstration of fitted distribution on the top keywords and to complement the results from chapter 5, we present the results in the form of table 8.1. According to this table for all the top 40 keywords exponential distribution and Gaussian distribution always have the least amount of difference from the mean values of the real data and they share the first rank, also the majority of rank 3 for these differences belongs to the t distribution 8.2(22 out of 40 keywords). Unexpectedly, the lower differences in mean values do not result in the significance of such variables, the set of [migration, innovation, implementation, machine learning, depression, stability, genus, estimation, type 1 diabetes, breast cancer, children] among these have a p-value lower than 0.05 and set of [migration, implementation, parameter estimation, machine learning] does not follow a normal distribution. For the closeness of differences in standard deviation, the majority of the first rank belongs to the chi-squared distribution(15 out of 40), and the second rank is the exponential distribution. Regarding the closeness of standard deviation, an interesting point that can be observed is that no distribution shares the top ranks.

Table 8.1: The distributions of keywords and the difference between mean and standard deviation [part 1](lower rank suggests a smaller difference between fitted distribution and real data)

Keywords	1_rank_mean	2_rank_mean	1_rank_sd	2_rank_sd	1_rank_mean_p_val	2_rank_mean_p_val	1_rank_sd_p_val	2_rank_sd_p_val	1_rank_mean_dist	2_rank_mean_dist	1_rank_sd_dist	2_rank_sd_dist
sweden	1.5	1.5	1	2	0.066499	0.09149	0.048731	0.066499	expon	norm	chi2	expon
gender	1.5	1.5	1	2	0.906949	0.333795	0.772808	0.506099	expon	norm	weibull_min	gamma
sverige	1.5	1.5	1	2	0.149003	0.149125	0.095804	0.193433	expon	norm	chi2	weibull_min
heart failure	1.5	1.5	1	2	0.302557	0.095581	0.02471	0.302557	expon	norm	beta	expon
children	1.5	1.5	1	2	0.044477	0.223427	0.114617	0.034574	expon	norm	weibull_min	chi2
system identification	1.5	1.5	1	2	0.080422	0.065209	0.080422	0.016122	expon	norm	expon	beta
depression	1.5	1.5	1	2	0.00506	0.071031	0.00506	0.071031	expon	norm	chi2	norm
optimization	1.5	1.5	1	2	0.348167	0.107183	0.26024	0.000208	expon	norm	chi2	gamma
education	1.5	1.5	1	2	0.266565	0.318833	0.188689	0.266565	expon	norm	chi2	expon
quality of life	1.5	1.5	1	2	0.559402	0.718793	0.468702	0.000993	expon	norm	weibull_min	chi2
epidemiology	1.5	1.5	1	2	0.106672	0.223377	0.114617	0.16273	expon	norm	chi2	t
identification	1.5	1.5	1	2	0.545588	0.09902	0.260159	0.743565	expon	norm	chi2	weibull_min
simulation	1.5	1.5	1	2	0.111847	0.307198	0.067481	0.202451	expon	norm	weibull_min	chi2
covid-19	1.5	1.5	1	2	0.745011	0.977166	0.745011	0.90625	expon	norm	expon	chi2
inflammation	1.5	1.5	1	2	0.333135	0.697624	0.001002	0.333135	expon	norm	chi2	expon
implementation	1.5	1.5	1	2	0.002873	0.003379	0.003083	0.003083	expon	norm	beta	chi2
migration	1.5	1.5	1	2	0.000116	0.002688	0.000116	0.000116	expon	norm	weibull_min	chi2
communication	1.5	1.5	1	2	0.136067	0.185902	0.152929	0.136067	expon	norm	weibull_min	expon
estimation	1.5	1.5	1	2	0.018473	0.373259	2.40E-05	0.000111	expon	norm	gamma	chi2
stability	1.5	1.5	1	2	0.010674	0.052371	0.010674	0.010674	expon	norm	chi2	expon
learning	1.5	1.5	1	2	0.270103	0.258039	0.021491	0.270103	expon	norm	chi2	expon
prognosis	1.5	1.5	1	2	0.18062	0.849703	0.18062	0.929357	expon	norm	expon	beta
mortality	1.5	1.5	1	2	0.265559	0.073146	0.131729	0.411561	expon	norm	t	chi2
energy efficiency	1.5	1.5	1	2	0.071785	0.192479	0.025488	0.071785	expon	norm	gamma	expon
genus	1.5	1.5	1	2	0.018094	0.056727	0.018094	0.018094	expon	norm	chi2	beta
innovation	1.5	1.5	1	2	0.002585	0.086021	0.002585	0.002585	expon	norm	weibull_min	chi2
sustainability	1.5	1.5	1	2	0.24721	0.88193	0.24721	0.482288	expon	norm	expon	beta
type 1 diabetes	1.5	1.5	1	2	0.026319	0.617698	0.026319	0.703831	expon	norm	expon	rayleigh
internet	1.5	1.5	1	2	0.104126	0.330297	0.717344	0.104126	expon	norm	weibull_min	expon
rehabilitation	1.5	1.5	1	2	0.68727	0.314881	0.800787	0.080748	expon	norm	weibull_min	chi2
apoptosis	1.5	1.5	1	2	0.089876	0.369135	0.089876	0.484602	expon	norm	expon	gamma
visualization	1.5	1.5	1	2	0.354835	0.564677	0.354835	0.342451	expon	norm	expon	beta
parameter estimation	1.5	1.5	1	2	0.117545	0.016373	0.040565	2.46E-06	expon	norm	chi2	gamma
machine learning	1.5	1.5	1	2	0.004494	0.038716	0.004495	0.004494	expon	norm	chi2	beta
stress	1.5	1.5	1	2	0.330467	0.800145	0.568553	0.568553	expon	norm	weibull_min	chi2
adolescents	1.5	1.5	1	2	0.606371	0.35587	0.061002	0.606371	expon	norm	chi2	expon
breast cancer	1.5	1.5	1	2	0.042949	0.66248	0.042949	0.978131	expon	norm	expon	gamma
pain	1.5	1.5	1	2	0.591297	0.713627	0.591297	0.713622	expon	norm	expon	t
dementia	1.5	1.5	1	2	0.496436	0.20858	0.797017	0.463496	expon	norm	chi2	weibull_min
pregnancy	1.5	1.5	1	2	0.311279	0.556253	0.438903	0.311279	expon	norm	weibull_min	expon

Table 8.2: The distributions of keywords and the difference between mean and standard deviation [part 2](lower rank suggests a smaller difference between fitted distribution and real data)

Keywords	3_rank_mean	4_rank_mean	3_rank_sd	4_rank_sd	3_rank_mean_p_val	4_rank_mean_p_val	3_rank_sd_p_val	4_rank_sd_p_val	3_rank_mean_dist	4_rank_mean_dist	3_rank_sd_dist	4_rank_sd_dist
sweden	3	4	3	4	0.091491	0.047733	0.000754	0.091491	t	rayleigh	gamma	t
gender	3	4	3	4	0.506099	0.445258	0.000754	0.091491	gamma	rayleigh	gamma	t
sverige	3	4	3	4	0.073917	0.193433	0.433274	0.335033	rayleigh	weibull_min	t	chi2
heart failure	3	4	3	4	0.095582	0.059105	0.433274	0.335033	t	rayleigh	t	chi2
children	3	4	3	4	0.223412	0.097695	0.193433	0.087325	t	rayleigh	beta	t
system identification	3	4	3	4	0.366387	0.366428	0.193433	0.087325	chi2	gamma	beta	t
depression	3	4	3	4	0.030888	0.00506	0.02729	0.095581	rayleigh	chi2	gamma	norm
optimization	3	4	3	4	0.107195	0.26024	0.02729	0.095581	t	chi2	gamma	norm
education	3	4	3	4	0.188689	0.026177	0.044477	0.223412	chi2	weibull_min	expon	t
quality of life	3	4	3	4	0.602827	0.749821	0.044477	0.223412	rayleigh	t	expon	t
epidemiology	3	4	3	4	0.161399	0.114617	0.508309	0.366428	rayleigh	chi2	t	gamma
identification	3	4	3	4	0.099016	0.140271	0.508309	0.366428	t	rayleigh	t	gamma
simulation	3	4	3	4	0.307201	0.273574	0.00506	0.00506	t	rayleigh	beta	expon
covid-19	3	4	3	4	0.977166	0.992108	0.00506	0.00506	t	rayleigh	beta	expon
inflammation	3	4	3	4	0.697628	0.768676	0.348167	0.107183	t	weibull_min	expon	norm
implementation	3	4	3	4	0.003083	0.000978	0.348167	0.107183	chi2	rayleigh	expon	norm
migration	3	4	3	4	0.000426	0.000116	0.026177	0.318833	rayleigh	chi2	weibull_min	norm
communication	3	4	3	4	0.185936	0.081533	0.026177	0.318833	t	rayleigh	weibull_min	norm
estimation	3	4	3	4	0.373259	0.159464	0.001077	0.559402	t	weibull_min	gamma	expon
stability	3	4	3	4	0.052371	0.010674	0.001077	0.559402	t	chi2	gamma	expon
learning	3	4	3	4	0.258037	0.021491	0.223377	0.106672	t	chi2	norm	expon
prognosis	3	4	3	4	0.849705	0.847021	0.223377	0.106672	t	gamma	norm	expon
mortality	3	4	3	4	0.411561	0.300574	0.000755	0.545588	chi2	beta	gamma	expon
energy efficiency	3	4	3	4	0.19248	0.071785	0.000755	0.545588	t	chi2	gamma	expon
genus	3	4	3	4	0.013207	0.018094	0.111847	0.000743	rayleigh	chi2	expon	gamma
innovation	3	4	3	4	0.086027	0.002585	0.111847	0.000743	t	weibull_min	expon	gamma
sustainability	3	4	3	4	0.88193	0.99532	0.898569	0.977166	t	gamma	gamma	norm
type 1 diabetes	3	4	3	4	0.809384	0.809385	0.898569	0.977166	chi2	gamma	gamma	norm
internet	3	4	3	4	0.852413	0.852413	0.768676	0.658812	gamma	chi2	weibull_min	beta
rehabilitation	3	4	3	4	0.78025	0.402561	0.768676	0.658812	beta	rayleigh	weibull_min	beta
apoptosis	3	4	3	4	0.369132	0.484586	0.003379	0.000978	t	chi2	norm	rayleigh
visualization	3	4	3	4	0.642732	0.642771	0.003379	0.000978	chi2	gamma	norm	rayleigh
parameter estimation	3	4	3	4	0.040565	0.078079	0.000116	0.002688	chi2	weibull_min	beta	norm
machine learning	3	4	3	4	0.006133	0.004495	0.000116	0.002688	rayleigh	chi2	beta	norm
stress	3	4	3	4	0.800145	0.624573	0.185936	0.185902	t	rayleigh	t	norm
adolescents	3	4	3	4	0.355856	0.204597	0.185936	0.185902	t	rayleigh	t	norm
breast cancer	3	4	3	4	0.978129	0.66249	0.018473	0.159464	chi2	t	expon	weibull_min
pain	3	4	3	4	0.713622	0.608645	0.018473	0.159464	t	rayleigh	expon	weibull_min
dementia	3	4	3	4	0.208584	0.463496	0.052371	0.052371	t	weibull_min	t	norm
pregnancy	3	4	3	4	0.556258	0.608506	0.052371	0.052371	t	rayleigh	t	norm

8.2 Bayesian Network Visualization

The Bayesian networks which were described in 5, contain more than 200 nodes and the corresponding connections. In this section, the top 20 nodes that have the most connections in each graph were selected and visualized, and the corresponding adjacency matrix was given. To view the results at full resolution, please visit project repository.

8.2.1 Media and Information technology

Many of these keywords indicate a strong interdisciplinary nature, merging fields like computer science, biology, physics, and design. The custom-designed structure contains many visualization-related nodes, a subset of it is shown in the next page:

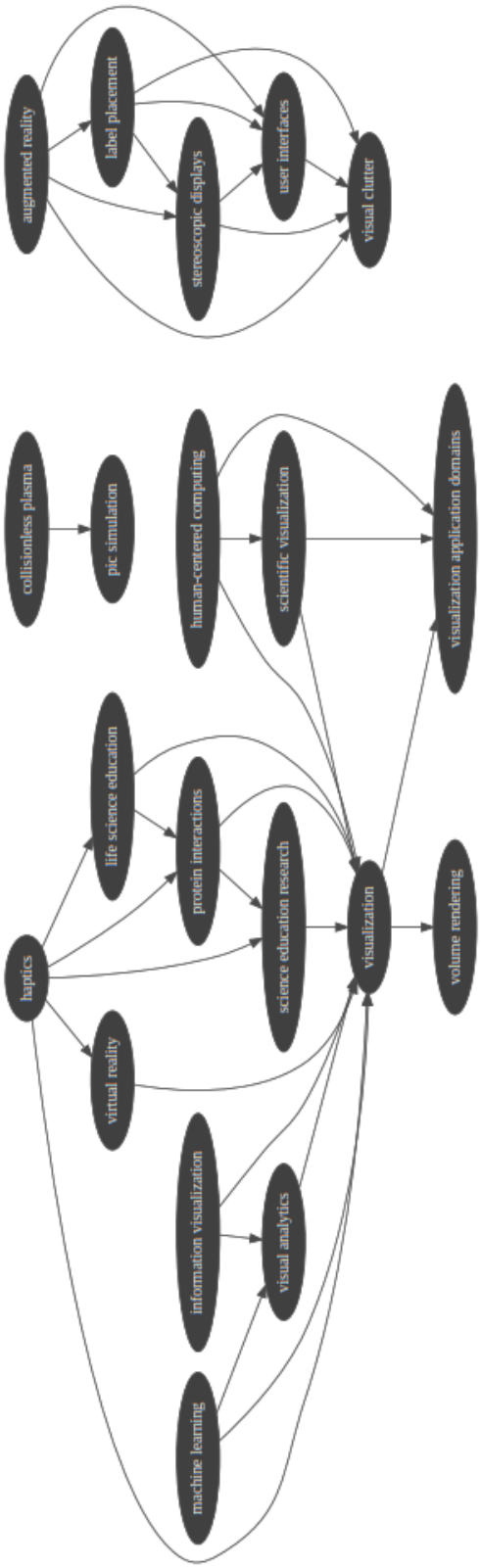


Figure 8.1: Network of top keywords for MIT dataset

for these nodes adjacency matrix representing the connections in the designed Bayesian network is given below:

Table 8.3: custom-designed adjacency matrix for the MIT dataset(Part 1)

	visualization	machine learning	augmented reality	haptics	information visualization	visual analytics
visualization	0	0	0	0	0	0
machine learning	1	0	0	0	0	1
augmented reality	0	0	0	0	0	0
haptics	1	0	0	0	0	0
information visualization	1	0	0	0	0	1
visual analytics	1	0	0	0	0	0
protein interactions	1	0	0	0	0	0
visual clutter	0	0	0	0	0	0
stereoscopic displays	0	0	0	0	0	0
human-centered computing	1	0	0	0	0	0
user interfaces	0	0	0	0	0	0
label placement	0	0	0	0	0	0
collisionless plasma	0	0	0	0	0	0
pic simulation	0	0	0	0	0	0
virtual reality	1	0	0	0	0	0
volume rendering	0	0	0	0	0	0
visualization application domains	0	0	0	0	0	0
science education research	1	0	0	0	0	0
scientific visualization	1	0	0	0	0	0
life science education	1	0	0	0	0	0

Table 8.4: custom-designed adjacency matrix for the MIT dataset(Part 2)

	protein interactions	visual clutter	stereoscopic displays	human-centered computing	user interfaces	label placement
visualization	0	0	0	0	0	0
machine learning	0	0	0	0	0	0
augmented reality	0	1	1	0	1	1
haptics	1	0	0	0	0	0
information visualization	0	0	0	0	0	0
visual analytics	0	0	0	0	0	0
protein interactions	0	0	0	0	0	0
visual clutter	0	0	0	0	0	0
stereoscopic displays	0	1	0	0	1	0
human-centered computing	0	0	0	0	0	0
user interfaces	0	1	0	0	0	0
label placement	0	1	1	0	1	0
collisionless plasma	0	0	0	0	0	0
pic simulation	0	0	0	0	0	0
virtual reality	0	0	0	0	0	0
volume rendering	0	0	0	0	0	0
visualization application domains	0	0	0	0	0	0
science education research	0	0	0	0	0	0
scientific visualization	0	0	0	0	0	0
life science education	1	0	0	0	0	0

Table 8.5: custom-designed adjacency matrix for the MIT dataset(Part 3)

	collisionless plasma	pic simulation	virtual reality	volume rendering	visualization application domains	science education research
visualization	0	0	0	1	1	0
machine learning	0	0	0	0	0	0
augmented reality	0	0	0	0	0	0
haptics	0	0	1	0	0	1
information visualization	0	0	0	0	0	0
visual analytics	0	0	0	0	0	0
protein interactions	0	0	0	0	0	1
visual clutter	0	0	0	0	0	0
stereoscopic displays	0	0	0	0	0	0
human-centered computing	0	0	0	0	1	0
user interfaces	0	0	0	0	0	0
label placement	0	0	0	0	0	0
collisionless plasma	0	1	0	0	0	0
pic simulation	0	0	0	0	0	0
virtual reality	0	0	0	0	0	0
volume rendering	0	0	0	0	0	0
visualization application domains	0	0	0	0	0	0
science education research	0	0	0	0	0	0
scientific visualization	0	0	0	0	1	0
life science education	0	0	0	0	0	0

Table 8.6: custom-designed adjacency matrix for the MIT dataset(Part 4)

	scientific visualization	life science education
visualization	0	0
machine learning	0	0
augmented reality	0	0
haptics	0	1
information visualization	0	0
visual analytics	0	0
protein interactions	0	0
visual clutter	0	0
stereoscopic displays	0	0
human-centered computing	1	0
user interfaces	0	0
label placement	0	0
collisionless plasma	0	0
pic simulation	0	0
virtual reality	0	0
volume rendering	0	0
visualization application domains	0	0
science education research	0	0
scientific visualization	0	0
life science education	0	0

The learned network structure contains:

Figure 8.2: Learned Network of top keywords for MIT dataset

and the adjacency matrix would be:

Table 8.7: Learned network adjacency matrix for the MIT dataset(Part 1)

	visualization	haptics	machine learning	information visualization	label placement	human-centered computing
visualization	0	1	0	0	0	1
haptics	0	0	0	0	0	0
machine learning	0	0	0	0	0	0
information visualization	0	0	0	0	0	0
label placement	0	0	0	0	0	0
human-centered computing	0	0	0	0	0	0
augmented reality	0	0	0	0	0	0
virtual reality	0	0	0	0	0	0
evolution	0	0	0	0	0	0
air traffic control	0	0	0	0	0	0
volume rendering	0	0	0	0	0	0
visual analytics	1	0	1	0	0	0
protein interactions	0	0	0	0	0	0
transfer function	0	0	0	0	0	0
collisionless plasma	0	0	0	0	0	0
technology education	0	0	0	0	0	0
text visualization	0	0	0	0	0	0
pic simulation	0	0	0	0	0	0
categorical data	0	0	0	0	0	0
automation	0	0	0	0	0	0

Table 8.8: Learned network adjacency matrix for the MIT dataset(Part 2)

	augmented reality	virtual reality	evolution	air traffic control	volume rendering	visual analytics
visualization	0	0	1	0	0	0
haptics	0	1	0	0	0	0
machine learning	0	0	0	0	0	0
information visualization	0	0	0	0	0	0
label placement	1	0	0	1	0	0
human-centered computing	0	0	0	0	0	0
augmented reality	0	0	0	0	0	0
virtual reality	0	0	0	0	0	0
evolution	0	0	0	0	0	0
air traffic control	0	0	0	0	0	0
volume rendering	0	0	0	0	0	0
visual analytics	0	0	0	0	0	0
protein interactions	0	0	0	0	0	0
transfer function	0	0	0	0	0	0
collisionless plasma	0	0	0	0	0	0
technology education	0	0	0	0	0	0
text visualization	0	0	0	0	0	1
pic simulation	0	0	0	0	0	0
categorical data	0	0	0	0	0	0
automation	0	0	0	0	0	0

Table 8.9: Learned network adjacency matrix for the MIT dataset(Part 3)

	protein interactions	transfer function	collisionless plasma	technology education	text visualization	pic simulation
visualization	0	0	0	0	0	0
haptics	1	0	0	0	0	0
machine learning	0	0	0	0	0	0
information visualization	0	0	0	0	1	0
label placement	0	0	0	0	0	0
human-centered computing	0	0	0	0	0	0
augmented reality	0	0	0	0	0	0
virtual reality	0	0	0	0	0	0
evolution	0	0	0	0	0	0
air traffic control	0	0	0	0	0	0
volume rendering	0	1	0	0	0	0
visual analytics	0	0	0	0	0	0
protein interactions	0	0	0	0	0	0
transfer function	0	0	0	0	0	0
collisionless plasma	0	0	0	0	0	1
technology education	0	0	0	0	0	0
text visualization	0	0	0	0	0	0
pic simulation	0	0	0	0	0	0
categorical data	0	0	0	0	0	0
automation	0	0	0	0	0	0

Table 8.10: Learned network adjacency matrix for the MIT dataset(Part 4)

	categorical data	automation
visualization	0	0
haptics	0	0
machine learning	0	0
information visualization	1	0
label placement	0	0
human-centered computing	0	0
augmented reality	0	0
virtual reality	0	0
evolution	0	0
air traffic control	0	1
volume rendering	0	0
visual analytics	0	0
protein interactions	0	0
transfer function	0	0
collisionless plasma	0	0
technology education	0	0
text visualization	0	0
pic simulation	0	0
categorical data	0	0
automation	0	0

8.2.2 Computer and Information science

This dataset reflects a diverse range of topics within computer science and engineering such as embedded systems, AI, modeling, and safety-critical systems. The top nodes:

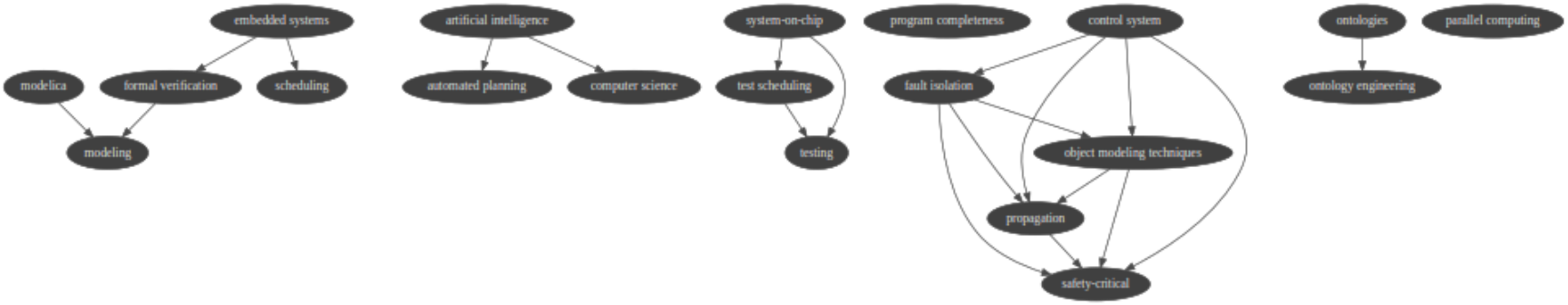


Figure 8.3: Network of top keywords for IDA dataset

and the adjacency matrix:

Table 8.11: custom-designed adjacency matrix for the IDA dataset(Part 1)

	embedded systems	system-on-chip	modeling	fault isolation	ontology engineering	parallel computing
embedded systems	0	0	0	0	0	0
system-on-chip	0	0	0	0	0	0
modeling	0	0	0	0	0	0
fault isolation	0	0	0	0	0	0
ontology engineering	0	0	0	0	0	0
parallel computing	0	0	0	0	0	0
propagation	0	0	0	0	0	0
testing	0	0	0	0	0	0
modelica	0	0	1	0	0	0
formal verification	0	0	1	0	0	0
control system	0	0	0	1	0	0
artificial intelligence	0	0	0	0	0	0
safety-critical	0	0	0	0	0	0
object modeling techniques	0	0	0	0	0	0
program completeness	0	0	0	0	0	0
scheduling	0	0	0	0	0	0
ontologies	0	0	0	0	1	0
computer science	0	0	0	0	0	0
test scheduling	0	0	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.12: custom-designed adjacency matrix for the IDA dataset(Part 2)

	propagation	testing	modelica	formal verification	control system	artificial intelligence
embedded systems	0	0	0	1	0	0
system-on-chip	0	1	0	0	0	0
modeling	0	0	0	0	0	0
fault isolation	1	0	0	0	0	0
ontology engineering	0	0	0	0	0	0
parallel computing	0	0	0	0	0	0
propagation	0	0	0	0	0	0
testing	0	0	0	0	0	0
modelica	0	0	0	0	0	0
formal verification	0	0	0	0	0	0
control system	1	0	0	0	0	0
artificial intelligence	0	0	0	0	0	0
safety-critical	0	0	0	0	0	0
object modeling techniques	1	0	0	0	0	0
program completeness	0	0	0	0	0	0
scheduling	0	0	0	0	0	0
ontologies	0	0	0	0	0	0
computer science	0	0	0	0	0	0
test scheduling	0	1	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.13: custom-designed adjacency matrix for the IDA dataset(Part 3)

	safety-critical	object modeling techniques	program completeness	scheduling	ontologies	computer science
embedded systems	0	0	0	1	0	0
system-on-chip	0	0	0	0	0	0
modeling	0	0	0	0	0	0
fault isolation	1	1	0	0	0	0
ontology engineering	0	0	0	0	0	0
parallel computing	0	0	0	0	0	0
propagation	1	0	0	0	0	0
testing	0	0	0	0	0	0
modelica	0	0	0	0	0	0
formal verification	0	0	0	0	0	0
control system	1	1	0	0	0	0
artificial intelligence	0	0	0	0	0	1
safety-critical	0	0	0	0	0	0
object modeling techniques	1	0	0	0	0	0
program completeness	0	0	0	0	0	0
scheduling	0	0	0	0	0	0
ontologies	0	0	0	0	0	0
computer science	0	0	0	0	0	0
test scheduling	0	0	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.14: custom-designed adjacency matrix for the IDA dataset(Part 4)

	test scheduling	automated planning
embedded systems	0	0
system-on-chip	1	0
modeling	0	0
fault isolation	0	0
ontology engineering	0	0
parallel computing	0	0
propagation	0	0
testing	0	0
modelica	0	0
formal verification	0	0
control system	0	0
artificial intelligence	0	1
safety-critical	0	0
object modeling techniques	0	0
program completeness	0	0
scheduling	0	0
ontologies	0	0
computer science	0	0
test scheduling	0	0
automated planning	0	0

similarly for the learned network:

Figure 8.4: Learned Network of top keywords for IDA dataset

The adjacency matrix of the learned network is computed as:

Table 8.15: Learned network adjacency matrix for the IDA dataset(Part 1)

	scheduling	parallel computing	testing	embedded systems	fault isolation	propagation
scheduling	0	0	0	0	0	0
parallel computing	1	0	0	0	0	0
testing	0	0	0	1	0	0
embedded systems	1	0	0	0	0	0
fault isolation	0	0	0	0	0	1
propagation	0	0	0	0	0	0
system-on-chip	0	0	1	0	0	0
modelica	0	0	0	0	0	0
test scheduling	0	0	1	0	0	0
safety-critical	0	0	0	0	0	1
knowledge representation	0	0	0	0	0	0
object modeling techniques	0	0	0	0	0	1
ontology engineering	0	0	0	0	0	0
gpu	0	1	0	0	0	0
test access mechanism	0	0	0	0	0	0
ontologies	0	0	0	0	0	0
computer science	0	0	0	0	0	0
modeling	0	0	0	0	0	0
formal verification	0	0	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.16: Learned network adjacency matrix for the IDA dataset(Part 2)

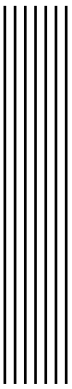
	system-on-chip	modelica	test scheduling	safety-critical	knowledge representation	object modeling techniques
scheduling	0	0	0	0	0	0
parallel computing	0	0	0	0	0	0
testing	0	0	0	0	0	0
embedded systems	0	0	0	0	0	0
fault isolation	0	0	0	1	0	1
propagation	0	0	0	0	0	0
system-on-chip	0	0	1	0	0	0
modelica	0	0	0	0	0	0
test scheduling	0	0	0	0	0	0
safety-critical	0	0	0	0	0	1
knowledge representation	0	0	0	0	0	0
object modeling techniques	0	0	0	0	0	0
ontology engineering	0	0	0	0	0	0
gpu	0	0	0	0	0	0
test access mechanism	0	0	0	0	0	0
ontologies	0	0	0	0	0	0
computer science	0	0	0	0	0	0
modeling	0	0	0	0	0	0
formal verification	0	0	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.17: Learned network adjacency matrix for the IDA dataset(Part 3)

	ontology engineering	gpu	test access mechanism	ontologies	computer science	modeling
scheduling	0	0	0	0	0	0
parallel computing	0	0	0	0	0	0
testing	0	0	0	0	0	0
embedded systems	0	0	0	0	0	0
fault isolation	0	0	0	0	0	0
propagation	0	0	0	0	0	0
system-on-chip	0	0	1	0	0	0
modelica	0	0	0	0	0	0
test scheduling	0	0	1	0	0	0
safety-critical	0	0	0	0	0	0
knowledge representation	0	0	0	1	0	0
object modeling techniques	0	0	0	0	0	0
ontology engineering	0	0	0	0	0	0
gpu	0	0	0	0	0	0
test access mechanism	0	0	0	0	0	0
ontologies	1	0	0	0	0	0
computer science	0	0	0	0	0	0
modeling	0	0	0	0	0	0
formal verification	0	0	0	0	0	0
automated planning	0	0	0	0	0	0

Table 8.18: Learned network adjacency matrix for the IDA dataset(Part 4)

	formal verification	automated planning
scheduling	0	0
parallel computing	0	0
testing	0	0
embedded systems	1	0
fault isolation	0	0
propagation	0	0
system-on-chip	0	0
modelica	0	0
test scheduling	0	0
safety-critical	0	0
knowledge representation	0	0
object modeling techniques	0	0
ontology engineering	0	0
gpu	0	0
test access mechanism	0	0
ontologies	0	0
computer science	0	0
modeling	0	0
formal verification	0	0
automated planning	0	0



Bibliography

[1] Khushnood Abbas, Mohammad Kamrul Hasan, Alireza Abbasi, Umi Asma' Mokhtar, Altaf Khan, Norul Huda, Shi Dong, Shayla Islam, Dabiah Alboaneen, and Fatima Rayan. "Predicting the future popularity of academic publications using deep learning by considering it as temporal citation networks". In: *IEEE Access* 11 (Jan. 2023), pp. 83052–83068. DOI: 10.1109/access.2023.3290906.

[2] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Giovanni Felici. "Predicting publication long-term impact through a combination of early citations and journal impact factor". In: *Journal of Informetrics* 13 (Feb. 2019), pp. 32–49. DOI: 10.1016/j.joi.2018.11.003.

[3] Ali Abrishami and Sadegh Aliakbary. "Predicting citation counts based on deep neural network learning techniques". In: *Journal of Informetrics* 13 (May 2019), pp. 485–499. DOI: 10.1016/j.joi.2019.02.011.

[4] Alberto Arteta Albert, Luis Fernando de Mingo López, and Nuria Gómez Blas. "Multilinear Weighted Regression (MWE) with Neural Networks for trend prediction". In: *Applied Soft Computing* 82 (Sept. 2019), p. 105555. DOI: 10.1016/j.asoc.2019.105555.

[5] Hanan Aljuaid, Rimsha Iftikhar, Shahbaz Ahmad, Muhammad Asif, and Tanvir Afzal. "Important citation identification using sentiment analysis of in-text citations". In: *Telematics and Informatics* 56 (Sept. 2020), p. 101492. DOI: 10.1016/j.tele.2020.101492.

[6] Chainarong Amornbunchornvej, Elena Zheleva, and Tanya Y Berger-Wolf. "Variable-Lag Granger Causality for Time Series Analysis". In: *arXiv (Cornell University)* (Oct. 2019). DOI: 10.1109/dsaa.2019.00016.

[7] Paul Baker. "Querying keywords". In: *Journal of English Linguistics* 32 (Dec. 2004), pp. 346–359. DOI: 10.1177/0075424204269894.

[8] Ozge Baris-Tuzemen and Johan Lyhagen. "Revisiting the role of climate change on crop production: evidence from Mediterranean countries". In: *Environment, development and sustainability* (May 2024). DOI: 10.1007/s10668-024-04991-x.

[9] Krzysztof Bartoszek. "Simulating an infinite mean waiting time". In: *Mathematica Applicanda* 47 (July 2019). DOI: 10.14708/ma.v47i1.6476.

[10] Juliane Begenau, Maryam Farboodi, and Laura Veldkamp. "Big data in finance and the growth of large firms". In: *Journal of Monetary Economics* 97 (Aug. 2018), pp. 71–87. DOI: 10.1016/j.jmoneco.2018.05.013. URL: https://www.sciencedirect.com/science/article/pii/S0304393218302174.

[11] Concha Bielza and Pedro Larrañaga. "Bayesian networks in neuroscience: a survey". In: *Frontiers in Computational Neuroscience* 8 (Oct. 2014). DOI: 10.3389/fncom.2014.00131. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4199264/.

[12] David Blei, Blei@cs Edu, Andrew Ng, Michael Jordan, and Jordan@cs Edu. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com.

[13] Katy Börner, Chaomei Chen, and Kevin W Boyack. "Visualizing knowledge domains". In: *Annual Review of Information Science and Technology* 37 (Jan. 2005), pp. 179–255. DOI: 10.1002/aris.1440370106.

[14] Lutz Bornmann. "Alternative metrics in scientometrics: a meta-analysis of research into three altmetrics". In: *Scientometrics* 103 (Mar. 2015), pp. 1123–1144. DOI: 10.1007/s11192-015-1565-y.

[15] Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. "Topic-based document segmentation with probabilistic latent semantic analysis". In: (Nov. 2002). DOI: 10.1145/584792.584829.

[16] Laura E Brown, Ioannis Tsamardinos, and Constantin F Aliferis. "A novel algorithm for scalable and accurate Bayesian network learning." In: *PubMed* 107 (Jan. 2004), pp. 711–5.

[17] , Andrés Cano, Javier G Castellano, and Serafín Moral. "Combining gene expression data and prior knowledge for inferring gene regulatory networks via Bayesian networks using structural restrictions". In: *Statistical applications in genetics and molecular biology* 18 (May 2019). DOI: 10.1515/sagmb-2018-0042.

[18] Chaomei Chen. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature". In: *Journal of the American Society for Information Science and Technology* 57 (2006), pp. 359–377. DOI: 10.1002/asi.20317.

[19] Gang Chen, Daniel R Glen, Ziad S Saad, J. Paul Hamilton, Moriah E Thomason, Ian H Gotlib, and Robert W Cox. "Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis". In: *Computers in Biology and Medicine* 41 (Dec. 2011), pp. 1142–1155. DOI: 10.1016/j.compbiomed.2011.09.004.

[20] Wen-Ta Chiu and Yuh-Shan Ho. "Bibliometric analysis of tsunami research". In: *Scientometrics* 73 (July 2007), pp. 3–17. DOI: 10.1007/s11192-005-1523-1.

[21] Chua Chongyong and Hongchoon Ong. "SCIENCE & TECHNOLOGY Comparison of Scoring Functions on Greedy Search Bayesian Network Learning Algorithms". In: *Pertanika Journal of Science & Technology* 25 (2017).

[22] Lawrence J. Christiano. "Christopher A. Sims and Vector Autoregressions*". In: *The Scandinavian Journal of Economics* 114 (Nov. 2012), pp. 1082–1104. DOI: 10.1111/j.1467-9442.2012.01737.x. URL: http://pages.stern.nyu.edu/~dbackus/Identification/Christiano_on_Sims_SJE_12.pdf.

[23] Diego Colombo and Marloes Maathuis. "Order-Independent Constraint-Based Causal Structure Learning". In: *Journal of Machine Learning Research* 15 (2014), pp. 3921–3962. URL: https://www.jmlr.org/papers/volume15/colombo14a/colombo14a.pdf.

[24] Muge Deng. "Time Series Analysis of China's Air Passenger Traffic Amid the COVID-19 Pandemic". In: *BCP Business & Management* 34 (Dec. 2022), pp. 1168–1178. DOI: 10.54691/bcpbm.v34i.3155.

[25] *Department of Computer and Information Science (IDA)*. liu.se. URL: https://liu.se/en/organisation/liu/ida.

[26] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. "How to conduct a bibliometric analysis: An overview and guidelines". In: *Journal of Business Research* 133 (Sept. 2021), pp. 285–296. DOI: 10.1016/j.jbusres.2021.04.070.

[27] *Elsevier Developer Portal*. dev.elsevier.com. URL: https://dev.elsevier.com/.

[28] Pär Emanuelson. "Performance enhancement in a well-structured pattern matcher through partial evaluation". In: (Jan. 1980).

[29] Nicholas E. Evangelopoulos. "Latent semantic analysis". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 4 (Aug. 2013), pp. 683–692. DOI: 10.1002/wcs.1254.

[30] ROBERT A FAIRTHORNE. "Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction". In: *Journal of Documentation* 25 (Apr. 1969), pp. 319–343. DOI: 10.1108/eb026481.

[31] Paolo Federico, Florian Heimerl, Steffen Koch, and Silvia Miksch. "A survey on visual approaches for analyzing scientific literature and patents". In: *IEEE Transactions on Visualization and Computer Graphics* 23 (Sept. 2017), pp. 2179–2198. DOI: 10.1109/tvcg.2016.2610422.

[32] Meng Di-fei, Liu Na, Li Ming-xia, and Su Hao-long. "An Improved Dynamic Collaborative Filtering Algorithm Based on LDA". In: *IEEE Access* (2021), pp. 1–1. DOI: 10.1109/access.2021.3094519.

[33] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. "Using Bayesian networks to analyze expression data". In: (Apr. 2000). DOI: 10.1145/332306.332355.

[34] Jonathan Furner. "The Ethics of Evaluative Bibliometrics". In: *The MIT Press eBooks* (Jan. 2014). DOI: 10.7551/mitpress/9445.003.0008.

[35] Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S Yu. "A survey of parallel sequential pattern mining". In: *ACM Transactions on Knowledge Discovery from Data* 13 (July 2019), pp. 1–34. DOI: 10.1145/3314107.

[36] *A languagebased approach to measuring scholarly impact*. MACHINE LEARNINIGINTERNATIONAL WORKSHOP THEN CONFERENCE. Madison, Wis.; [International Machine Learning Society]; c, Jan. 2010, pp. 375–382.

[37] Domenico Giannone, Michele Lenza, and Giorgio E. Primiceri. "Prior Selection for Vector Autoregressions". In: *Review of Economics and Statistics* 97 (May 2015), pp. 436–451. DOI: 10.1162/rest_a_00483.

[38] Wolfgang Glänzel and Pei-Shan Chi. "Scientometrics 2.0 – and beyond? Background, promises, challenges and limitations". In: 2016. URL: https://api.semanticscholar.org/CorpusID:157949856.

[39] Amanda H Goodall. "Should top universities be led by top researchers and are they?" In: *Journal of Documentation* 62 (May 2006), pp. 388–411. DOI: 10.1108/00220410610666529. URL: https://libweb.anglia.ac.uk/referencing/files/QuickHarvardGuide2018.pdf.

[40] T. L. Griffiths and M. Steyvers. "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101 (Feb. 2004), pp. 5228–5235. DOI: 10.1073/pnas.0307752101.

[41] Arzu Tugce Guler, Cathelijn J. F Waaijer, and Magnus Palmblad. "Scientific workflows for bibliometrics". In: *Scientometrics* 107 (Feb. 2016), pp. 385–398. DOI: 10.1007/s11192-016-1885-6.

[42] Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, and Thoudam Doren Singh. "Prediction of Research Trends using LDA based Topic Modeling". In: *Global Transitions Proceedings* (Apr. 2022). DOI: 10.1016/j.gltp.2022.03.015.

[43] Scott Hacker and Abdulnasser Hatemi-J. "A bootstrap test for causality with endogenous lag length choice: theory and application in finance". In: *Journal of Economic Studies* 39 (May 2012), pp. 144–160. DOI: 10.1108/01443581211222635.

[44] Sture Hägglund. "Contributions to the development of methods and tools for interactive design of applications software". In: (Jan. 1980).

[45] David Heckerman. "Bayesian Networks for Data Mining". In: *Data Mining and Knowledge Discovery* 1 (1997), pp. 79–119. DOI: 10.1023/a:1009730122752.

[46] Donna L Hoffman and Morris B Holbrook. "The intellectual structure of consumer research: A bibliometric study of author cocitations in the first 15 years of the journal of consumer research". In: *Journal of Consumer Research* 19 (Mar. 1993), p. 505. DOI: 10.1086/209319.

[47] Matthew Hoffman, Francis Bach, and David Blei. *Online Learning for Latent Dirichlet Allocation*. Neural Information Processing Systems, 2010. URL: https://proceedings.neurips.cc/paper/3902-online-learning-for-latentdirichlet-allocation.

[48] Thomas Hofmann. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99* (1999). DOI: 10.1145/312624.312649. URL: https://dl.acm.org/citation.cfm?id=312649.

[49] Linlin Hou, Ji Zhang, Ou Wu, Ting Yu, Zhen Wang, Zhao Li, Jianliang Gao, Yingchun Ye, and Rujing Yao. "Method and dataset entity mining in scientific literature: A CNN + BiLSTM model with self-attention". In: *Knowledge-Based Systems* 235 (Jan. 2022), p. 107621. DOI: 10.1016/j.knosys.2021.107621.

[50] Zhiping Hou, Fasheng Cui, Yongheng Meng, Tonghui Lian, and Caihua Yu. "Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis". In: *Tourism Management* 74 (Oct. 2019), pp. 276–289. DOI: 10.1016/j.tourman.2019.03.009.

[51] Beibei Hu, Yang Ding, Xianlei Dong, Yi Bu, and Ying Ding. "On the relationship between download and citation counts: An introduction of Granger-causality inference". In: *Journal of Informetrics* 15 (May 2021), p. 101125. DOI: 10.1016/j.joi.2020.101125.

[52] Zhaosong Huang, Daniel Witschard, Kostiantyn Kucher, and Andreas Kerren. "VA + embeddings STAR: A State-of-the-Art report on the use of embeddings in visual analytics". In: *Computer Graphics Forum* 42 (June 2023), pp. 539–571. DOI: 10.1111/cgf.14859.

[53] Angela Hullmann and Martin Meyer. "Publications and patents in nanotechnology". In: *Scientometrics* 58 (2003), pp. 507–527. DOI: 10.1023/b:scie.0000006877.45467.a7.

[54] Gogtay N J and Thatte U M. "Principles of correlation analysis". In: *The Journal of the Association of Physicians of India* 65 (Mar. 2017), pp. 78–81. URL: https://pubmed.ncbi.nlm.nih.gov/28462548/.

[55] RAHUL JHA, AMJAD-ABU JBARA, VAHED QAZVINIAN, and DRAGOMIR R RADEV. "NLP-driven citation analysis for scientometrics". In: *Natural Language Engineering* 23 (Jan. 2016), pp. 93–130. DOI: 10.1017/s1351324915000443.

[56] Guan Jiancheng and Wang Junxia. "Evaluation and interpretation of knowledge production efficiency". In: *Scientometrics* 59 (2004), pp. 131–155. DOI: 10.1023/b:scie.0000013303.25298.ae.

[57] Erland Jungert. "Synthesizing database structures from a user oriented data model". In: (Jan. 1980).

[58] Johan Källström and Fredrik Heintz. "Learning Agents for Improved Efficiency and Effectiveness in Simulation-Based Training". In: (Jan. 2020), pp. 1–2.

[59] Victoria Kayser and Knut Blind. "Extending the knowledge base of foresight: The contribution of text mining". In: *Technological Forecasting and Social Change* 116 (Mar. 2017), pp. 208–215. DOI: 10.1016/j.techfore.2016.10.017.

[60] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. "A survey of Bayesian Network structure learning". In: *Artificial Intelligence Review* (Jan. 2023). DOI: 10.1007/s10462-022-10351-w.

[61] Kostiantyn Kucher, Carita Paradis, and Andreas Kerren. "The state of the art in sentiment visualization". In: *Computer Graphics Forum* 37 (June 2017), pp. 71–96. DOI: 10.1111/cgf.13217.

[62] *An appraisal of publication embedding techniques in the context of conventional bibliometric relatedness measures*. 18th International Conference on Scientometrics and Informetrics, ISSI 2021. 2021, pp. 633–638. URL: https://www.scopus.com/inward/record.uri?eid=2s2.085112659182&partnerID=40&md5=56f4e1c8050d1160406b347acba46a6a.

[63] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis". In: *Discourse Processes* 25 (Jan. 1998), pp. 259–284. DOI: 10.1080/01638539809545028.

[64] Helge Langseth and Luigi Portinale. "Bayesian networks in reliability". In: *Reliability Engineering & System Safety* 92 (Jan. 2007), pp. 92–108. DOI: 10.1016/j.ress.2005.11.037.

[65] P. Larranaga, M. Poza, Y. Yurramendi, R.H. Murga, and C.M.H. Kuijpers. "Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (1996), pp. 912–926. DOI: 10.1109/34.537345.

[66] Colin I.S.G Lee, Will Felps, and Yehuda Baruch. "Toward a taxonomy of career studies through bibliometric visualization". In: *Journal of Vocational Behavior* 85 (Dec. 2014), pp. 339–351. DOI: 10.1016/j.jvb.2014.08.008.

[67] J LEE, W CHUNG, and E KIM. "Structure Learning of Bayesian Networks Using Dual Genetic Algorithm". In: *IEICE transactions on information and systems* E91-D (Jan. 2008), pp. 32–43. DOI: 10.1093/ietisy/e91-d.1.32.

[68] Woo Hyoung Lee. "How to identify emerging research fields using scientometrics: An example in the field of Information Security". In: *Scientometrics* 76 (July 2008), pp. 503–525. DOI: 10.1007/s11192-007-1898-2.

[69] Lei Lei, Yaochen Deng, and Dilin Liu. "Examining research topics with a dependency-based noun phrase extraction method: a case in accounting". In: *Library Hi Tech* ahead-of-print (2023). DOI: 10.1108/lht-12-2019-0247.

[70] Xuqing Liu and Xinsheng Liu. "Structure learning of Bayesian networks by continuous particle swarm optimization algorithms". In: *Statistical computation and simulation/Journal of statistical computation and simulation* 88 (Feb. 2018), pp. 1528–1556. DOI: 10.1080/00949655.2018.1440395.

[71] Wei Lu, Shengzhi Huang, Jinqing Yang, Yi Bu, Qikai Cheng, and Yong Huang. "Detecting research topic trends by author-defined keyword frequency". In: *Information Processing & Management* 58 (July 2021), p. 102594. DOI: 10.1016/j.ipm.2021.102594.

[72] H Madsen, D Lawrence, M Lang, M Martinkova, and T.R Kjeldsen. "Review of trend analysis and climate change projections of extreme precipitation and floods in Europe". In: *Journal of Hydrology* 519 (Nov. 2014), pp. 3634–3650. DOI: 10.1016/j.jhydrol.2014.11.003.

[73] Dimitris Margaritis and Sebastian Thrun. *Bayesian Network Induction via Local Neighborhoods*. Neural Information Processing Systems, 1999. URL: https://proceedings.neurips.cc/paper/1999/hash/5d79099fcdf499f12b79770834c0164a-Abstract.html.

[74] Noah Mauthe, Ulf Kargen, and Nahid Shahmehri. "A Large-Scale Empirical Study of Android App Decompilation". In: (Mar. 2021). DOI: 10.1109/saner50967.2021.00044.

[75] Tim Mazzarol, Thierry Volery, Noelle Doss, and Vicki Thein. "Factors influencing small business start-ups". In: *International Journal of Entrepreneurial Behavior & Research* 5 (Apr. 1999), pp. 48–63. DOI: 10.1108/13552559910274499.

[76] *Media and Information Technology*. liu.se. URL: https://liu.se/en/research/media-and-information-technology-mit.

[77] McWilliams J Michael, Ellen Meara, Alan M Zaslavsky, and John Z Ayanian. "Differences in control of cardiovascular disease and diabetes by race, ethnicity, and education: U.S. trends from 1999 to 2006 and effects of medicare coverage". In: *Annals of Internal Medicine* 150 (Apr. 2009), pp. 505–515. DOI: 10.7326/0003-4819-150-8-200904210-00005. URL: https://pubmed.ncbi.nlm.nih.gov/19380852/.

[78] Theophano Mitsa. *Temporal data mining*. CRC Press, Mar. 2010. DOI: 10.1201/9781420089776.

[79] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Feb. 2021. URL: https://books.google.com/books?hl=en&lr=&id=tCIgEAAAQBAJ&oi=fnd&pg=PR13&dq=inIntroduction+to+Linear+Regression+Analysis&ots=lgsdYpf0Ln&sig=ZVNZhIMVzRg9QmZIycDmmnRSoAA.

[80] Luca Mora, Mark Deakin, and Alasdair Reid. "Combining co-citation clustering and text-based analysis to reveal the main development paths of smart cities". In: *Technological Forecasting and Social Change* 142 (May 2019), pp. 56–69. DOI: 10.1016/j.techfore.2018.07.019.

[81] Peter Moser. "The impact of legislative institutions on public policy: a survey". In: *European Journal of Political Economy* 15 (Mar. 1999), pp. 1–33. DOI: 10.1016/s0176-2680(98)00038-x.

[82] Jeanette A. Mumford and Joseph D. Ramsey. "Bayesian networks for fMRI: A primer". In: *NeuroImage* 86 (Feb. 2014), pp. 573–582. DOI: 10.1016/j.neuroimage.2013.10.020.

[83] Shanthala Nagaraja and Kiran Yarehalli Chandrappa. "Topic Modelling". In: *International Journal of Innovative Technology and Exploring Engineering* 9 (Dec. 2019), pp. 482–485. DOI: 10.35940/ijitee.b1124.1292s19.

[84] Chris J. Needham, James R. Bradford, Andrew J. Bulpitt, and David R. Westhead. "A Primer on Learning in Bayesian Networks for Computational Biology". In: *PLoS Computational Biology* 3 (2007), e129. DOI: 10.1371/journal.pcbi.0030129.

[85] Magdalena Osińska. "On the Interpretation of Causality in Granger's Sense". In: *Dynamic Econometric Models* 11 (Dec. 2011). DOI: 10.12775/dem.2011.009.

[86] Denis Luiz Marcello Owa. "Identification of Topics from Scientific Papers through Topic Modeling". In: *Open Journal of Applied Sciences* 10 (2021), pp. 541–548. DOI: 10.4236/ojapps.2021.104038.

[87] Shweta Pandey, Neeraj Pandey, and Deepak Chawla. "Market segmentation based on customer experience dimensions extracted from online reviews using data mining". In: *Journal of Consumer Marketing* (July 2023). DOI: 10.1108/jcm-10-2022-5654.

[88] Judea Pearl. "An introduction to causal inference". In: *The International Journal of Biostatistics* 6 (Jan. 2010). DOI: 10.2202/1557-4679.1203.

[89] T Penfield, Baker M J, R Scoble, and Wykes M C. "Assessment, evaluations, and definitions of research impact: A review". In: *Research Evaluation* 23 (Oct. 2013), pp. 21–32. DOI: 10.1093/reseval/rvt021. URL: https://academic.oup.com/rev/article/23/1/21/2889056.

[90] Bryan Pesta, John Fuerst, and. "Bibliometric keyword analysis across seventeen years (2000–2016) of intelligence articles". In: *Journal of Intelligence* 6 (Oct. 2018), p. 46. DOI: 10.3390/jintelligence6040046. URL: https://doaj.org/article/8ddb6c98aa8b4b0a85e5fb837420f6fe.

[91] John W Pratt and Jean D Gibbons. "Kolmogorov-Smirnov Two-Sample Tests". In: *Springer series in statistics* (Jan. 1981), pp. 318–344. DOI: 10.1007/978-1-4612-5931-2_7.

[92] Junping Qiu, Zhao Rong, Siluo Yang, and Ke Dong. *Informetrics*. Springer Singapore, Jan. 2017. DOI: 10.1007/978-981-10-4032-0.

[93] Radim Řehůřek and Petr Sojka. *Software Framework for Topic Modelling with Large Corpora*. University of Malta, May 2010. URL: https://repozitar.cz/publication/15725/?lang=cs.

[94] Paulo Rita and Ricardo F Ramos. "Global research trends in consumer behavior and sustainability in e-commerce: A bibliometric analysis of the knowledge structure". In: *Sustainability* 14 (Aug. 2022), p. 9455. DOI: 10.3390/su14159455.

[95] Guido Van Rossum. *Python reference manual*. Iuniverse.Com, Inc, 2000.

[96] Imran Hussain Shah, Charlie Hiles, and Bruce Morley. "How do oil prices, macroeconomic factors and policies affect the market for renewable energy?" In: *Applied Energy* 215 (Apr. 2018), pp. 87–97. DOI: 10.1016/j.apenergy.2018.01.084.

[97] Ali Shojaie and Emily B. Fox. "Granger Causality: A Review and Recent Advances". In: *Annual Review of Statistics and Its Application* 9 (Nov. 2021). DOI: 10.1146/annurev-statistics-040120-010930.

[98] Peter Spirtes and Clark Glymour. "An Algorithm for Fast Recovery of Sparse Causal Graphs". In: *Social Science Computer Review* 9 (Apr. 1991), pp. 62–72. DOI: 10.1177/089443939100900106.

[99] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Jan. 2001. URL: https://books.google.com/books?hl=en&lr=&id=OZ0vEAAAQBAJ&oi=fnd&pg=PR9&dq=Causation.

[100] Jörgen I Stenlund, Konrad J Schönborn, and Gunnar E Höst. "Design and validation of a deep evolutionary time visual instrument (DET-Vis)". In: *Evolution* 15 (July 2022). DOI: 10.1186/s12052-022-00170-6.

[101] Jennifer Ann Stevenson and Jin Zhang. "A temporal analysis of institutional repository research". In: *Scientometrics* 105 (Sept. 2015), pp. 1491–1525. DOI: 10.1007/s11192-015-1728-x.

[102] William J Sutherland, David Goulson, Simon G Potts, and Lynn V Dicks. "Quantifying the impact and relevance of scientific research". In: *PLoS ONE* 6 (Nov. 2011). Ed. by Tammy Clifford, e27537. DOI: 10.1371/journal.pone.0027537.

[103] Shaheen Syed and Marco Spruit. "FullText or abstract? Examining topic coherence scores using latent dirichlet allocation". In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA. IEEE, Jan. 2017. DOI: 10.1109/dsaa.2017.61.

[104] Yee Teh, David Newman, and Max Welling. *A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation*. Neural Information Processing Systems, 2006. URL: https://proceedings.neurips.cc/paper/2006/hash/532b7cbe070a3579f424988a040752f2-Abstract.html.

[105] Tung Thanh Nguyen, Tho Thanh Quan, and Tuoi Thi Phan. "Sentiment search: an emerging trend on social media monitoring systems". In: *Aslib Journal of Information Management* 66 (Sept. 2014). Ed. by Alexander and Fran Dr Ulrike Spree, pp. 553–580. DOI: 10.1108/ajim-12-2013-0141.

[106] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. "The max-min hill-climbing Bayesian network structure learning algorithm". In: *Machine Learning* 65 (Mar. 2006), pp. 31–78. DOI: 10.1007/s10994-006-6889-7.

[107] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. "Text mining techniques for patent analysis". In: *Information Processing & Management* 43 (Sept. 2007), pp. 1216–1247. DOI: 10.1016/j.ipm.2006.11.011.

[108] Robert R Tucci. "Introduction to Judea Pearl's Do-Calculus". In: *arXiv (Cornell University)* (Jan. 2013). DOI: 10.48550/arxiv.1305.5506.

[109] Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. "Learning causal networks with latent variables from multivariate information in genomic data". In: *PLOS Computational Biology* 13 (Oct. 2017). Ed. by Jennifer Listgarten, e1005662. DOI: 10.1371/journal.pcbi.1005662.

[110] Johan A Wallin. "Bibliometric methods: Pitfalls and possibilities". In: *Basic Clinical Pharmacology Toxicology* 97 (Nov. 2005), pp. 261–275. DOI: 10.1111/j.1742-7843.2005.pto_139.x.

[111] Jian Wang, Reinhilde Veugelers, and Paula Stephan. "Bias against novelty in science: A cautionary tale for users of bibliometric indicators". In: *Research Policy* 46 (Oct. 2017), pp. 1416–1436. DOI: 10.1016/j.respol.2017.06.006. URL: https://www.sciencedirect.com/science/article/abs/pii/S0048733317301038.

[112] Jun Wang and Klaus Mueller. "The Visual Causality Analyst: An Interactive Interface for Causal Reasoning". In: *IEEE Transactions on Visualization and Computer Graphics* 22 (Jan. 2016), pp. 230–239. DOI: 10.1109/tvcg.2015.2467931.

[113] Mengyang Wang and Lihe Chai. "Three new bibliometric indicators/approaches derived from keyword analysis". In: *Scientometrics* 116 (May 2018), pp. 721–750. DOI: 10.1007/s11192-018-2768-9.

[114] P. Weber, G. Medina-Oliva, C. Simon, and B. Iung. "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas". In: *Engineering Applications of Artificial Intelligence* 25 (June 2012), pp. 671–682. DOI: 10.1016/j.engappai.2010.06.002.

[115] Howard D White and Katherine W McCain. "Visualizing a discipline: An author co-citation analysis of information science, 1972–1995". In: *Journal of the American Society for Information Science* 49 (1998), pp. 327–355. DOI: 10.1002/(sici)1097-4571(19980401)49:4<327::aid-asi4>3.0.co;2-4.

[116] D. J. Wilkinson. "Bayesian methods in bioinformatics and computational systems biology". In: *Briefings in Bioinformatics* 8 (Dec. 2006), pp. 109–116. DOI: 10.1093/bib/bbm007.

[117] Richard Williams and Lutz Bornmann. "Sampling issues in bibliometric analysis". In: *Journal of Informetrics* 10 (Nov. 2016), pp. 1225–1232. DOI: 10.1016/j.joi.2015.11.004.

[118] Alyson G. Wilson and Aparna V. Huzurbazar. "Bayesian networks for multilevel system reliability". In: *Reliability Engineering & System Safety* 92 (Oct. 2007), pp. 1413–1420. DOI: 10.1016/j.ress.2006.09.003.

[119] Erjia Yan and Ying Ding. "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other". In: *Journal of the American Society for Information Science and Technology* 63 (May 2012), pp. 1313–1326. DOI: 10.1002/asi.22680.

[120] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. "A survey on causal inference". In: *ACM Transactions on Knowledge Discovery from Data* 15 (Oct. 2021), pp. 1–46. DOI: 10.1145/3444944.

[121] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. "Clustering scientific documents with topic modeling". In: *Scientometrics* 100 (May 2014), pp. 767–786. DOI: 10.1007/s11192-014-1321-8.

[122] Bin Yu, Jia-Meng Xu, Shan Li, Cheng Chen, Rui-Xin Chen, Lei Wang, Yan Zhang, and Ming-Hui Wang. "Inference of time-delayed gene regulatory networks based on dynamic Bayesian network hybrid learning method". In: *Oncotarget* 8 (Sept. 2017), pp. 80373–80392. DOI: 10.18632/oncotarget.21268.

[123] Tian Yu, Guang Yu, Peng-Yu Li, and Liang Wang. "Citation impact prediction for scientific papers using stepwise regression analysis". In: *Scientometrics* 101 (Mar. 2014), pp. 1233–1252. DOI: 10.1007/s11192-014-1279-6.

[124] Ruizhi Zhang, Jian Wang, and Yajun Mei. "Search for evergreens in science: A functional data analysis". In: *Journal of Informetrics* 11 (Aug. 2017), pp. 629–644. DOI: 10.1016/j.joi.2017.05.007.

[125] Xianglu Zhou and Haiyang Jia. "Computational Methods for Inferring and Reconstructing Gene Regulatory Networks". In: *Applied and computational engineering* 13 (Oct. 2023), pp. 229–239. DOI: 10.54254/2755-2721/13/20230738.

[126] Ke Zhu and Hanzhong Liu. "Confidence intervals for parameters in high-dimensional sparse vector autoregression". In: *Computational Statistics & Data Analysis* (Oct. 2021), p. 107383. DOI: 10.1016/j.csda.2021.107383.

[127] Shuang Zhu, Huan Meng, Zhanjun Gu, and Yuliang Zhao. "Research trend of nanoscience and nanotechnology – A bibliometric analysis of Nano Today". In: *Nano Today* 39 (Aug. 2021), p. 101233. DOI: 10.1016/j.nantod.2021.101233.

[128] Zikai Zhu, Peng Su, Sean Zhong, Jiayu Huang, Suranjan Ottikkutti, Kaveh Nazem Tahmasebi, Zhuo Zou, Lirong Zheng, and Dejiu Chen. "Using a VAE-SOM architecture for anomaly detection of flexible sensors in limb prosthesis". In: *Journal of industrial information integration* 35 (Oct. 2023), pp. 100490–100490. DOI: 10.1016/j.jii.2023.100490.