Differences between journal and conference in computer science: a bibliometric view based on Bayesian network

Mingyue Sun\textsuperscript{1,2}, Mingliang Yue\textsuperscript{1,2}, Tingcan Ma\textsuperscript{1,2†}

\textsuperscript{1}Wuhan Library, Chinese Academy of Sciences, China
\textsuperscript{2}Department of Information Resources Management, University of Chinese Academy of Sciences, China

Abstract

Purpose: This paper aims to investigate the differences between conference papers and journal papers in the field of computer science based on Bayesian network.

Design/methodology/approach: This paper investigated the differences between conference papers and journal papers in the field of computer science based on Bayesian network, a knowledge-representative framework that can model relationships among all variables in the network. We defined the variables required for Bayesian networks modeling, calculated the values of each variable based Aminer dataset (a literature data set in the field of computer science), learned the Bayesian network and derived some findings based on network inference.

Findings: The study found that conferences are more attractive to senior scholars, the academic impact of conference papers is slightly higher than journal papers, and it is uncertain whether conference papers are more innovative than journal papers.

Research limitations: The study was limited to the field of computer science and employed Aminer dataset as the sample. Further studies involving more diverse datasets and different fields could provide a more complete picture of the matter.

Practical implications: By demonstrating that Bayesian networks can effectively analyze issues in Scientometrics, the study offers valuable insights that may enhance researchers’ understanding of the differences between journal and conference in computer science.

Originality/value: Academic conferences play a crucial role in facilitating scholarly exchange and knowledge dissemination within the field of computer science. Several studies

\footnote{Corresponding author: Tingcan Ma (matc@whlib.ac.cn; ORCID: 0000-0001-5985-384X)}

Citation: Sun, M. Y., Yue, M. L., Ma, T. C.. (2023). Differences between journal and conference in computer science: a bibliometrics view based on Bayesian network. Journal of Data and Information Science, 8(3), 47–60. https://doi.org/10.2478/jdis-2023-0017

Received: Apr. 28, 2023
Accepted: May 22, 2023
have been conducted to examine the distinctions between conference papers and journal papers in terms of various factors, such as authors, citations, h-index and others. Those studies were carried out from different (independent) perspectives, lacking a systematic examination of the connections and interactions between multiple perspectives. This paper supplements this deficiency based on Bayesian network modeling.

**Keywords:** Conference papers; Journal papers; Computer science; Bibliometrics; Bayesian network

1 Introduction

Compared with other disciplines, conferences are very important in computer science (Freyne et al., 2010) and they have become the main channel for scientific research and dissemination in the field (Shamir, 2010). Due to the rapid pace of technological innovation in computer science, conferences are particularly suitable for researchers to communicate their findings in a timely manner (Fortnow, 2009). The importance of conferences has prompted scholars to consider the differences between conferences and journals in the field.

Up to now, various factors that differentiate computer science conferences and journals have been studied. For instance, scholars have analyzed the relationship between publication type (journal, conference) and citation count (Birman & Schneider, 2009; Eckmann et al., 2012; Fernández Izquierdo et al., 2007; Freyne et al., 2010; Qian et al., 2017; Vrettas & Sanderson, 2015; Wainer et al., 2011), publication Ranking (CCF A B C) and citation count (Freyne et al., 2010; Qian et al., 2017; Vrettas & Sanderson, 2015), authorship and citation count (Qian et al., 2017), publication type and authors (Kumari & Kumar, 2020), and publication type and institutions (Kumari & Kumar, 2020). Their research is beneficial for a deeper understanding of the differences between computer conferences and journals from various perspectives. However, these studies were carried out from different (independent) perspectives, lacking a systematic examination of the connections and interactions between multiple perspectives.

Recently, Sun et al. (2023) have made progress in modeling and analyzing the relationships among citation and influencing factors using Bayesian network (BN) in a systematic manner. In the paper, 20 factors that are related to paper citation have been modeled by BN so that the relationships among citation and the influencing factors are concisely represented (based on the network structure and parameter of the BN), and the interactions among them are dynamically recognized (based on the network reasoning).

Therefore, we investigated the differences between conference papers and journal papers in the field of computer science based on Bayesian network from the
Differences between journal and conference in computer science: a bibliometric view based on Bayesian network

perspective of systematic interaction among multiple factors. We defined the variables required for Bayesian networks (BN) modeling, including variables corresponding to publication types and CCF classification indicators that have been newly added. Then, we calculated the values and states of each variable from more than 5 million paper records based Aminer dataset (a literature data set in the field of computer science). At last, we analyzed the characteristics of conferences and journals from different perspectives, compared our findings with existing conclusions, resulting in some interesting findings.

The remainder of this paper is as follows. Section 2 gives some preliminary knowledge of Bayesian network and the network construction method proposed by Sun et al. (2023). Section 3 gives the Bayesian network construction process. Section 4 shows some findings based on the inference of the Bayesian network. Section 5 concludes the paper.

2 Preliminary knowledge

A Bayesian network (Pearl, 1988) is defined as a pair \((G, P)\). \(G = (V, E)\) is a Directed Acyclic Graph (DAG) used to capture the structure of the knowledge domain, \(V = \{X_1, X_2, \ldots, X_n\}\) is a set of nodes given by the random variables of the domain, \(E \subseteq V \times V\) is a set of directed edges representing the probabilistic conditional (in)dependencies among the nodes, a node \(X_i\) is a parent of another node \(X_j\) if there is an arc from \(X_i\) to \(X_j\). \(P\) is a set of parameters characterizing the joint probability distribution over \(X_1\) to \(X_n\), denoted as \(P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \pi(X_i))\), where \(\pi(X_i)\) represents the set of parents of \(X_i\) in \(G\).

The first task of using Bayesian network for knowledge analysis is to determine the network structure and parameters from data samples. One of the most popular approaches for Bayesian structure learning from data is based on searching and scoring. The aim is to search for the network structure that maximizes the scoring function defined to represent how well a structure fits a given data set. Among those methods, K2 is one of the most classic and commonly used methods.

Given a set of variables and data samples, the K2 algorithm starts from a graph with no edge, a number indicating maximum parents, and an order of variables, and adds an edge to the graph if the inclusion of the edge improves the scoring function most compared with other potential edges. The process repeats until no addition of a new edge can improve the score or all the maximum admissible parents are achieved. The K2 algorithm uses marginal likelihood as a score function. For a given variable, only preceding variables in the variable order can be considered as potential parents. The number of maximum parents is used to guarantee a concise representation of the domain knowledge.
It can be seen that K2 algorithm needs a variable order (that implies expert knowledge of the underlying domain) and try to find a good enough network structure based on the scoring function. However, in certain situations, there may not be a strict order among variables. In Sun et al. (2023), the method is extended to solve the problem of no strict orders among variables. The method is called the amended K2 algorithm, which is given in Algorithm 1.

Algorithm 1: amended K2 algorithm

```
input: the set of data samples \( D \), order of variable sets \( O \),
maximum number of parents \( m \)
output: a Bayesian Network \( G = (V, E, P) \)  
\( P \) is the conditional Probability table

1. \( E \leftarrow \emptyset, G \leftarrow (V, E) \);
2. \( S_m \leftarrow \) calculate the structure score of \( G \) based on \( D \);
3. repeat
   4. Subprocedure: find potential edge set \( E_p \);
   5. for each \( v_i \) in \( E_p \) do
   6. \( E_t \leftarrow E \cup E_i, G_t \leftarrow G_i \);
   7. \( S_i \leftarrow \) calculate the structure score of \( G_i \) based on \( D \);
   8. if \( S_i > S_m \) then
   9. \( S_m \leftarrow S_i, G \leftarrow G_i \);
10. until \( S_m \) not changed;
11. \( P \leftarrow \) evaluate \( P \) based on \( G \) and \( D \) using Maximum Likelihood Estimation;
12. return \( G = (V, E, P) \)
13. Subprocedure: find potential edge set \( E_p \)
14. \( E \leftarrow \emptyset \);
15. for each variable \( v_i \) in \( V \) do
16. if the number of parents of \( v_i \) is less than \( m \) then
   17. for each variable \( v_j \) other than \( v_i \) do
   18. if \( v_j \) is in the preorder set of \( v_i \) or \( v_i \) and \( v_j \) are in the same set and \( v_j \) is not a descendant of \( v_i \) then
   19. \( E \leftarrow E \cup (v_j, v_i) \);
20. return \( E \)
```

The input of Algorithm 1 includes the set of data samples \( D \), a number indicating the maximum number of parents of each variable (node) in the Bayesian network, and an order of variable sets \( O \). Compared with the classical K2 algorithm, the main modification of the amended K2 algorithm is included in the sub-procedure that finds the potential edge set. The edges satisfying the following conditions are regarded as potential legal edges: i) the ending node \( v_i \) has less than \( m \) parents, ii) the starting node \( v_j \) is in the preorder set with respect to \( v_i \), or \( v_i \) and \( v_j \) are in the same set and \( v_j \) is not a descendant of \( v_i \). The first condition is to guarantee a concise network (knowledge structure). The second condition is to guarantee all the potential networks are DAGs.
3 Method

In this section, we discuss factors (variables) specific to CS, then describe the underlying data set and the data processing procedure (used to calculate the factor values), and at last, show the learned Bayesian network.

3.1 Factors in computer science

We adopt the same set of factors given in Sun et al. (2023), which can be categorized into author-level, platform-level, internal, outcome, and influence factors. The internal factors are relevant to the paper itself, including novelty (pNov), disruption (pDisrupt), number of references (refNum), text readability (abRE), and text length (abLen). Author-related factors are relevant to the influence and collaboration level of the paper authors, including the number of published papers of the first author (pNumF) and of the author with maximum number of published papers (pNumM), total citations of the first author (tcF), total citations of the author with maximum total citations (tcM), h-index of the first author (HIF) and the maximum author (HIM), co-authorship network centrality degree of the first author (auCDF) and the maximum author (auCDM). Platform-related factors are relevant to the collaboration of the authors’ institutions, including the number of authors (auNum), number of institutes (instNum), cooperation network centrality degree of the first author’s institution (instCDF) and of the institution with the maximum cooperation network centrality (instCDM). The influence of a paper is measured with the Category Normalized Citation Impact (CNCI).

Further, since the aim of this paper is to investigate the distinctions between academic journals and conferences in CS, we include a Category indicator for each paper to identify its type (conference or journal). Besides, in view of the important influence of CCF Rankings\(^1\), also due to the conference papers do not have IF, we introduce CCF Rank as an alternative to JIFRank to signify the importance of the paper.

3.2 Data preprocessing

The Aminer dataset is used as the underlying data in this paper for the calculation of the factor values and BN learning (Tang, 2008). The Aminer dataset is a comprehensive collection of academic research papers and citation relationships and has been widely used in various research works relating to academic research evaluations (Abramo et al., 2019; Amjad et al., 2022; Shao et al., 2022; Song et al., 2018). The data set contains information related to 5,354,309 papers and 48,227,950

\(^1\)http://www.ccf.org.cn/sites/ccf/paiming.jsp.
citation relationships. It is one of the largest datasets available in computer science. The data set provides information on paper identification number (id), title, publication date (year), author details (including identification number (_id), name, institution name (org), and institution identification number (gid)), publication journal information (venue), abstract, citation count (n_citation), reference numbers (references), and complete citation relationships between papers. Based on the information, the factor values are calculated.

Except for Category and CCF Rank, before Bayesian network learning, factor values should be discretized into states. The value of Category can be J or C (for journal or conference) and of CCF Rank can be A B C. The discretization rules for other factors can be found in Table 1, which are given in Sun et al. (2023).

Table 1. Discretization rules of factors (Sun et al., 2023).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Discretization rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>pNov</td>
<td>0: zero; [0, 0.4]: low; (0.4, 0.6]: median; (0.6, 0.8]: mhigh; (0.8, 1]: high</td>
</tr>
<tr>
<td>pDisrupt</td>
<td>&lt;0: negzero; sort pDisrupt values and divide by top percentage interval: (70%, 100%]: low; (30%, 70%]: medium; (10%, 30%]: mhigh; (0%, 10%]: high</td>
</tr>
<tr>
<td>refNum</td>
<td>[0, 10]: ltTen; (10, 20]: tenTwenty; (20, 30]: twentyThirty; &gt;30: gtThirty</td>
</tr>
<tr>
<td>abRE</td>
<td>&gt;70: easy; (50, 70]: medium; (40, 50]: mhard; (30, 40]: hard; &lt;30: vhard (ref. Flesch, 1948)</td>
</tr>
<tr>
<td>abLen</td>
<td>&lt;600: short; (600, 800]: median; (800, 1000]: long; &gt;1000: vlong</td>
</tr>
<tr>
<td>pNumF</td>
<td>[0, 10]: ltTen; (10, 20]: tenTwenty; (20, 50]: twentyFifty; (50, 100]: FiftyHundred; &gt;100: gtHundred</td>
</tr>
<tr>
<td>pNumM</td>
<td>gtHundred</td>
</tr>
<tr>
<td>tcF</td>
<td>[0, 10]: ltTen; (10, 100]: tenHundred; (100, 500]: HundredFiveH; (500, 2000]: fiveHTwentyH; (2000, 10000]: twentyHHundredH; &gt;10000: gtHundredH</td>
</tr>
<tr>
<td>tcM</td>
<td>FiftyHundred</td>
</tr>
<tr>
<td>HIF</td>
<td>[0, 10]: ltTen; (10, 20]: tenTwenty; (20, 30]: twentyThirty; (30, 50]: thirtyFifty; (50, 100]:</td>
</tr>
<tr>
<td>HIM</td>
<td>FiftyHundred</td>
</tr>
<tr>
<td>auCDF</td>
<td>sort auCDF/auCDM values and divide by top percentage interval: (50%, 100%]: low; (20%, 50%]: mlow; (10%, 20%]: medium; (5%, 10%]: mhigh; (0%, 5%]: high</td>
</tr>
<tr>
<td>instCDF</td>
<td>high</td>
</tr>
<tr>
<td>instCDM</td>
<td></td>
</tr>
<tr>
<td>auNum</td>
<td>1: one; 2: tow; 3: three; 4: four; 5: five; &gt;5: gtfive</td>
</tr>
<tr>
<td>instNum</td>
<td></td>
</tr>
<tr>
<td>CNCI</td>
<td>(0, 0.3]: low; (0.3, 0.8]: mlow; (0.8, 1.2]: average; (1.2, 2]: mhigh; (2, 5]: vhigh; &gt;5: exhigh</td>
</tr>
</tbody>
</table>

3.3 Bayesian network construction

We learn Bayesian network structure and parameters based on the amended K2 algorithm proposed by Sun et al. (2023). Same as Sun et al. (2023), author-level factors, platform-level factors, internal factors, and influence factors are arranged in order as the input set order of the amended K2 algorithm. In this paper, Category and CCF Rank are considered as internal factors since they are paper-level factors. We
learn the BN based on the amended K2 algorithm and use Netica to visualize the learned Bayesian network, as shown in Figure 1.

\[\text{Figure 1. The learned Bayesian network.}\]

4 Findings

As shown in Figure 1, the initial state of the Bayesian network gives the marginal distribution of all the variables, which can be used to get a basic understanding of the field knowledge. For example, as shown in Figure 1, we observed that the \(\text{ltTen}\) pNumF probability is 68.5% and the \(\text{ltTen}\) HIF probability is 66.9%. If we use pNumF and HIF to indicate the academic impact of the first authors, the probabilities suggest that there are many junior scholars in CS. We also observed that the \(mhard\) and above abRE (the readability of abstracts) probability is more than 90%, indicating that abstracts of papers in CS are generally not easy to read.

By setting the values of certain variables, the network can acquire the conditional distribution of other variables based on network reasoning. Figure 2 gives an example of setting Category as the journal. It can be seen that the probability distribution of most variables has changed slightly. The fact means that, overall, there may not be a big difference between conferences and journals. For example, if we set Category as conference and journal respectively, the percentage of papers with more than three authors in conference papers (27.12%) is slightly higher than that in journal articles (25.66%), as reported in Franceschet (2011) and Kim (2019).
Therefore, we conduct a more fine-grained examination by setting the status of more variables, and try to find some interesting conclusions.

1. Conferences are more attractive to senior scholars

We use that pNumM and HIM indicate the academic impact of authors. As shown in Figure 3, when we change Category from journal to conference, the FiftyHundred and above pNumM probability is increased from 21.57% (12.9%+8.67%) to 32.2% (16.3%+15.9%), the FiftyHundred and above HIM probability is increased from 13.19% (8.42%+4.77%) to 14.64% (9.84%+4.8%), and the mhigh and above auCDM probability is increased from 1.43% (0.85%+0.58%) to 3.47% (1.94%+1.53%). The results show that the conference is more likely to attract senior scholars.
From the perspective of authors, as shown in Figure 4, we set the HIM as \textit{FiftyHundred}, pNumM as \textit{FiftyHundred}, and auCDM as \textit{mhigh}, the distribution of Category (conference, journal) is 68.6\%, 31.4\%. Similarly, by setting HIM as \textit{gtHundred}, pNumM as \textit{gtHundred}, and auCDM as \textit{high}, the distribution of Category (conference, journal) is 71.6\%, 28.4\%. It can be seen that senior scholars tend to prefer publishing papers in conference proceedings. The results further show that conferences are more attractive to senior scholars.

![Figure 4](http://www.jdis.org)

Figure 4. The distribution of Category by setting pNumM=(FiftyHundred, gtHundred), HIM=(FiftyHundred, gtHundred), auCDM=(mhigh, high).

(2) The academic impact (indicated by CNCI) of conference papers is slightly higher than journal papers

Although senior scholars generally prefer to publish papers at conferences, the Normalized Citation Impact (CNCI) of conference papers is not significantly higher than journal papers. As depicted in Figure 5(a), when changing the Category from conferences to journals, the probability of achieving a high CNCI score (mhigh or above) only slightly decreases from 11.90\%, 14.50\%, 6.13\% to 11\%, 13.4\%, 5.51\% respectively. We think this trend may be attributed to several reasons. First, the distribution of abstract length (abLen) changes from 22.60\%, 66\%, 10.60\%,0.82\% in conferences to 18.40\%, 55.80\%, 21.60\%, 4.21\% in journals, and the distribution of reference numbers (refNum) changes from 53\%, 31.2\%, 11\%, 4.81\% in conferences to 52.4\%, 31.1\%, 11.2\%, 5.23\% in journals, as demonstrated in Figure 6, suggesting that journal papers tend to provide more information (from the perspective of its own content and referenced knowledge). We think that more detailed content can bring (additional) impact to the journal articles.
Second, by considering CCF Rank, Figure 5(b) shows that when changing the Category from conference to journal and setting the CCF Rank as A, the probability of mhigh or above CNCI drops from 13.7%, 19.7%, 9.22% to 12%, 16.6%, 7.78%, respectively. These results indicate that the probability of mhigh or above CNCI is higher for grade A conferences than grade A journals. We repeated this process for Rank B and C and found that the probability of mhigh and above CNCI is slightly higher for grade B conferences (33.39% = 12.4% + 14.9% + 6.09%) than grade B journals (29.61% = 11.10% + 13.1% + 5.41%), and the probability of mhigh and above CNCI is lower in grade C conferences (24.90% = 10.10% + 10.70% + 4.10%) than in grade C journals (27.90% = 10.60% + 12.5% + 4.80%). The fact may indicate that higher quality (as indicated by CCF Rank) of a paper can result in more citations even if it is less informative. However, when the quality of a paper is relatively low, the richness of information it contains may have a greater impact on the citations it receives.

(3) It is uncertain whether conference papers are more innovative than journal
Considering that the rapid dissemination of research results is critical for computer science researchers, conference proceedings are generally published more quickly than traditional journals (Wainer et al., 2011). Due to this, people often feel that shorter publication cycles for conferences will prioritize the presentation of new research results at conferences, leading to the perception that conferences are more innovative. However, in Figure 7, we can see that when changing the Category from conference to journal, the probability of achieving \textit{mhigh} or above \textit{pNov} changes from 12.60\% and 18.70\% to 12.3\% and 19\% respectively, and the \textit{pDisrupt} changes from 21.3\% and 9.59\% to 21\% and 9.35\% respectively. This indicates that there is almost no difference in innovation and disruptiveness between conference papers and journal papers.

Further, if we consider both paper type and paper rank, the distributions of paper \textit{pNov} and \textit{pDisrupt} are also basically the same for different settings of Category and CCF Rank. As we can see in Figure 8, for various CCF ranks, the largest difference between the percentage of the \textit{mhigh} \textit{pNov} conference papers and journal papers is not more than 0.6\%, and of the high \textit{pNov} is not more than 0.7\%. Those difference for \textit{pDisrupt} is 0.2\% and 0.5\% for \textit{mhigh} and high respectively.

Based on the above results, we think it is uncertain whether conferences are more innovative than journals if innovation can be represented and measured by \textit{pNov} and \textit{pDisrupt}. However, as the most authoritative grade list in the field of computer science, the CCF rank has indeed guided researchers to submit their works with
higher-quality (often considered to be more innovative), at least in the authors’ own perception, to higher-level conferences or journals. The fact that there is no difference among pNov (and pDisrupt) over papers of different CCF ranks may also indicate that the applicability of the indicators in different disciplines and/or scenarios needs further discussion. That is, we think that both the problems of the innovation index (disruption index) itself and whether conferences are more innovative (indicated by certain evaluation index) still need further in-depth research.

Finally, we find that our findings are generally consistent with those given by Sun et al. (2023), which are drawn in the field of physics. For example, we also find that researchers have more influence on the impact of research works than institutes, and moderately innovative work can acquire more academic impact with respect to low-innovative and high-innovation work. The fact gives further evidence that Bayesian networks can be well applied for analyzing issues in Scientometrics.

5 Conclusion

In this paper, we investigated the differences between conference papers and journal papers in the field of computer science based on Bayesian network. We defined the variables required for Bayesian networks (BN) modeling, calculated the values and states of variables based Aminer dataset, and analyzed the characteristics of conferences and journals from different perspectives. We found that (1) conferences in the field of computer science are more attractive to senior scholars; (2) overall the academic impact (indicated by CNCI) of conference papers is slightly higher than journal papers, while the CCF C journal papers can get more impact than CCF C conference papers; (3) it is not certain whether conference papers are more innovative than journal papers.

We believe the results of this paper gave further evidence that Bayesian networks
Differences between journal and conference in computer science: a bibliometric view based on Bayesian network

can be well applied for analyzing issues in Scientometrics. Further work should be focused on refining the framework to be more sophistic, e.g., extending the variable set to include more related factors, incorporating more reliable expert knowledge (on variable order), introducing causal relationships, etc.

Author contributions

Mingyue Sun (summingyue22@mails.ucas.ac.cn): Methodology, Software, Data Curation, Visualization, Writing - Original Draft. Mingliang Yue (yueml@whlib.ac.cn): Conceptualization, Writing - review & editing, Funding acquisition. Tingcan Ma (matc@whlib.ac.cn): Conceptualization, Writing - review & editing, Funding acquisition.

Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Funding information

The work of this paper is supported by the Chinese Academy of Sciences Literature and Information capacity building project, Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2019176).

References


Copyright: © 2023 Mingyue Sun, Mingliang Yue, Tingcan Ma. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.