Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Automated Construction of Bayesian Networks of Chemical Process for Dynamic Risk Assessment

Kai Yin[ab], Hao Kang[c], Jinsong Zhao[a]*

[a] Tsinghua University, Department of Chemical Engineering, Beijing, 100084, China
[b] State Key Laboratory of Chemical Engineering and Low-carbon Technology, Engineering Department, Beijing, 100084, China
[c] Zhejiang Petroleum & Chemical Co., Ltd, Zhoushan, Zhejiang, China
* Corresponding Author: jinsongzhao@tsinghua.edu.cn.

ABSTRACT

Chemical processes are characterized by high complexity and inherent hazards, necessitating systematic and comprehensive risk assessment methodologies. However, traditional risk assessment approaches are often limited to static analyses and small-scale models, resulting in insufficient coverage and a lack of detailed, unit-level insights. This paper proposes an automated Bayesian network–based dynamic risk assessment (DRA) framework for full-process risk analysis in chemical plants. The proposed method automatically extracts knowledge from established risk analysis documents, such as HAZOP reports, and integrates it with dynamic process monitoring data, expert knowledge, and reliability databases. By effectively leveraging multi-source heterogeneous data, an equipment-level causal inference network is constructed automatically, addressing the high labor demand and limited scalability of conventional DRA approaches. The proposed framework is validated through an industrial FCC case study at a large petrochemical enterprise in Zhejiang, China. The results demonstrate that the method can effectively construct Bayesian causal networks for chemical processes and quantitatively evaluate both the likelihood and severity of accident consequences under various failure scenarios, thereby enabling systematic and dynamic risk assessment of chemical processes.

Keywords: dynamic risk assessment, Bayesian network, multi-source heterogeneous data, automated construction

1. INTRODUCTION

Chemical processes are inherently complex and fraught with various risks due to their dynamic nature and the involvement of hazardous materials. Traditionally, risk assessments in these environments have been static, relying on predefined assumptions and historical data. However, as industrial systems grow more complex and real-time monitoring becomes more prevalent, these conventional approaches are no longer sufficient to capture the dynamic behavior and evolving risks that can occur during operation.

Dynamic risk assessment (DRA) has emerged as a more suitable methodology to address these challenges. The main goal of DRA is to provide a real-time, adaptable framework for assessing the risks associated with chemical processes. This approach continuously updates risk levels based on variations in operating conditions, equipment degradation, human–machine interactions, and control system performance. By incorporating time-dependent factors, DRA offers a more comprehensive understanding of risk propagation and its evolution over time, enabling better decision-making in high-risk environments.

According to the definition by Khan et al.[1], DRA is a methodology that updates process risk estimates based on the performance of control systems, protection barriers, inspection and maintenance activities, human factors, and operational procedures. A key distinction between DRA and traditional risk assessment lies in the explicit incorporation of time-dependent factors. Importantly, risk analysis should quantify both the probability and the consequences of hazardous scenarios. Approaches that focus exclusively on occurrence probability without assessing consequence severity result in incomplete and potentially misleading risk evaluations.

In 2006, Meel and Seider[2] developed an integrated framework combining precursor information with real-time failure data to dynamically assess accident sequences, marking one of the earliest practical applications of DRA in chemical process safety.

Building upon tree-based models, subsequent research increasingly focused on dynamic risk assessment. Kalantarnia et al.[3] advanced DRA by integrating Bayesian failure modeling with consequence analysis. Their framework consists of two key modules: (i) an accident analysis module, which incorporates real-time process parameters into dynamic probability estimation using fault tree analysis; and (ii) a probability updating module, which iteratively updates probabilities through Bayesian inference as new data become available.

Between 2000 and 2009, Bayesian networks (BN) were widely adopted in DRA due to their natural compatibility with probabilistic updating and causal reasoning. Montani et al.[4] introduced dynamic Bayesian networks (DBN) and developed the DBNet tool to automatically convert dynamic fault trees into DBNs, representing a significant milestone in BN-based DRA. Kohda and Cui et al.[5] proposed a DBN-based dynamic reconfiguration method for safety monitoring systems, using time slices to model operational states and sensor reliability degradation. More recently, Shan et al.[6] employed fuzzy Bayesian networks (FBN) to quantify gas pipeline leakage risk, while Guo et al.[7] proposed a fuzzy dynamic Bayesian network (FDBN) framework for uncertainty propagation in DRA.

Machine learning methods have been increasingly applied to DRA. Adedigba et al.[8] used artificial neural networks (ANNs) for real-time accident probability prediction, while Wang et al.[9] applied long short-term memory (LSTM) networks for real-time risk forecasting. Other techniques, such as principal component analysis (PCA)[10] and support vector machines (SVM)[11], have also been explored.

Existing DRA methodologies mainly depend on Event Tree Analysis (ETA), Fault Tree Analysis (FTA), and Bow-Tie Analysis. Despite their effectiveness, most current approaches suffer from several limitations: incomplete risk characterization, limited network scale, and high reliance on expert knowledge. In particular, the construction of BN structures and the assignment of initial probabilities often require substantial manual effort, which hinders scalability and industrial applicability.

Although these studies collectively illustrate the evolution of DRA methodologies, most existing approaches still rely heavily on manual model construction and expert-driven probability assignment. This limitation is particularly pronounced for methods such as FBN, which are highly dependent on expert judgment. To address these challenges, this paper proposes an automated framework for extracting knowledge from multi-source heterogeneous data and constructing BN for DRA.

2. PRELIMINARIES

This section introduces the key techniques and algorithms employed in the proposed framework, including time-series Transformer-based causal discovery, sentence Transformer for heterogeneous textual data, and Bayesian parameter estimation. These techniques jointly enable the automated construction of a Bayesian causal network for DRA.

2.1 Time-Series Transformer-Baesd Causal Discovery

Causal discovery from process data plays a critical role in constructing the structural framework of a causal BN. In complex chemical process such as Fluid Catalytic Cracking (FCC), process variables are strongly coupled, dynamically evolving, and often influenced by delayed and indirect interactions. Traditional correlation-based methods are insufficient to capture such causal relationships.

In this work, causal discovery is performed using a time-series Transformer model[12] applied to DCS historical data. Transformer architectures, originally developed for sequence modeling, are particularly well suited for capturing long-range temporal dependencies and nonlinear interactions among multivariate time-series signals. By leveraging self-attention mechanisms, the model can learn how past values of one variable contribute to the future evolution of another.

Specifically, Granger causality is estimated within the Transformer framework: a variable X is considered to Granger-cause another variable Y if incorporating the historical information of X significantly improves the prediction of Y compared to using the history of Y alone. By evaluating such predictive contributions across all monitored variables, a directed causal graph is obtained, which serves as the initial structure of the BN. This data-driven approach enables scalable and automated identification of causal relationships in large-scale industrial systems without relying solely on expert-defined structures.

2.2 Sentence Transformer

To bridge the semantic gap between Bayesian network node descriptions expressed in unstructured natural language and reliability database entries provided in structured textual form, this study employs the Sentence-BERT (SBERT) framework for semantic vector representation. Conventional BERT models typically adopt a cross-encoder architecture for sentence similarity tasks, in which sentence pairs are jointly encoded. Although effective, this approach incurs high computational cost and is therefore unsuitable for large-scale database matching

and real-time retrieval. SBERT addresses this limitation by reformulating pretrained BERT models into a Siamese architecture with shared weights, allowing each sentence to be encoded independently into a fixed-dimensional embedding.

Given a Bayesian network node description A and a candidate database entry B, SBERT processes the two texts separately and generates contextualized token-level representations using BERT encoders. A pooling operation is then applied to aggregate token embeddings into a sentence-level vector. In this work, mean pooling is adopted, where the final sentence embedding u is computed as the average of all token embeddings:

$$u = \frac{1}{N}\sum_{i=1}^{N} h_i \quad (3)$$

where hi denotes the embedding of the ith token and N is the sentence length.

To ensure that the resulting embeddings exhibit meaningful geometric properties in the semantic space, SBERT is typically fine-tuned using a supervised classification objective. For a sentence pair (A, B), the corresponding embeddings u and v are concatenated with their absolute difference, and fed into a Softmax classifier trained on semantic entailment, contradiction, and neutrality labels. This training strategy encourages semantically similar sentences to be mapped closer together in the embedding space while pushing dissimilar sentences further apart.

During inference, Bayesian network node names and reliability database entries are mapped into high-dimensional vectors and their semantic correspondence is quantified using cosine similarity:

$$sim(A,B) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \quad (4)$$

The resulting similarity score ranges from [−1, 1], with higher values indicating stronger semantic similarity. This approach not only captures direct lexical overlap but also effectively identifies implicit semantic associations, such as those between instrumentation tags and their corresponding process semantics, thereby enabling robust matching between Bayesian network nodes and heterogeneous reliability data sources.

2.3 Bayesian Parameter Estimation

Parameter learning for conditional probability tables (CPTs) in BN is performed using Bayesian inference with Dirichlet priors. A commonly used baseline approach is maximum likelihood estimation (MLE), which estimates probabilities by simple frequency counting. For example, if node X has parent Y, and the dataset contains N instances where Y=High, among which X=Alarm occurs n times, then:

$$P(X = Alarm|Y = High) = \frac{n}{N} \quad (4)$$

However, MLE suffers from the zero-frequency problem. If a critical event has not occurred in the dataset, MLE assigns it a probability of zero. In process safety applications, “not observed” does not imply “impossible, ” and MLE completely ignores prior knowledge from reliability databases such as OREDA.

To overcome these limitations and integrate prior knowledge with observational data, Bayesian estimation is adopted. In discrete BN, node states follow a multinomial distribution, whose conjugate prior is the Dirichlet distribution. This approach is commonly referred to as Dirichlet smoothing.

A key parameter in Bayesian estimation is the Equivalent Sample Size (ESS), which quantifies the confidence placed in prior probabilities. ESS represents the number of virtual observations implied by the prior. Given a prior probability $P_{prior}(X_i = k|Pa_i = j)$, the corresponding pseudo-count is:

$$a_{ijk} = P_{prior}(X_i = k|Pa_i = j) \times ESS \quad (1)$$

The updated CPT is then computed as:

$$\theta_{ijk} = \frac{N_{ijk}+\alpha_{ijk}}{N_{ij}+\alpha_{ij}} \quad (2)$$

where N_{ijk} is the observed count from process data and α_{ij} is the sum of pseudo-counts. When large amounts of data are available, the influence of the prior diminishes. When data are scarce, the prior prevents zero-probability artifacts.

3. PROPOSED FRAMEWORK

This chapter provides a detailed description of the complete implementation workflow of the proposed automated dynamic risk assessment framework. The framework is designed to address two fundamental limitations of conventional risk assessment approaches: first, the subjectivity and high labor cost associated with manual construction of BN; and second, the inherent latency of static risk data, which fails to reflect the real-time operating state of chemical processes. As illustrated in Fig. 3, the implementation workflow is tightly integrated, starting from preprocessing of DCS data, proceeding through causal discovery and multi-source heterogeneous data extraction, and ultimately achieving the automated construction of a risk-oriented causal network for chemical processes.

3.1 Data Preprocessing and Transformer-Based Causal Discovery

The first stage of the framework aims to automatically extract the topology of the BN from the vast amount of Distributed Control System (DCS) data in chemical processes.

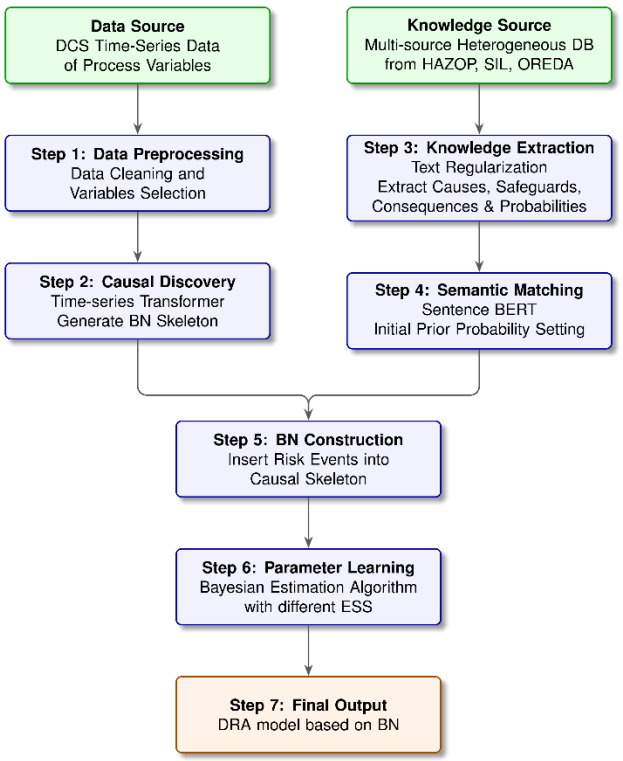

Figure 1. Framework of automated Bayesian causal network construction.

Initial preprocessing of the raw DCS time-series data involves selecting the key variables that have a critical impact on the safety of the process. In real-world chemical production plants, a single plant consists of thousands of instrument points that indicate the parameters of different systems and equipment. During the risk assessment process, some variables are closely related to the safe operation of the plant, while others are used to indicate production efficiency. In our risk analysis, we focus primarily on those variables that are tightly connected to production safety.

To extract these key variables, we collect a list of key variables from the HAZOP analysis table. The structure of the HAZOP analysis table is shown in Table 1. The HAZOP analysis report provides risk data on the occurrence probabilities of causes and consequences for a specific unit's risk scenarios, which are determined through brainstorming sessions by an expert panel. In the HAZOP analysis process, expert panels analyze all nodes in the P&ID diagrams and evaluate the risks in the process based on deviations observed in DCS data. We extract high-risk variables identified by the expert group during the HAZOP analysis, and then select these variables from the thousands of DCS data points for causal discovery.

After preprocessing, the study employs a Time-series Transformer model to capture generalized Granger causality relationships between process variables and construct the structural backbone of the BN. The Transformer architecture utilizes a self-attention mechanism to compute an attention weight matrix that quantifies the influence of historical states of variable j on the future state of variable i. This influence is expressed as the attention weight α_{ij}. By setting a significance threshold δ, if $\alpha_{ij}>\delta$, a causal link from node j to node i is determined. The algorithm iterates through all pairs of process variables, generating an adjacency matrix that describes the global causal relationships of the system. This matrix is directly transformed into a directed acyclic graph (DAG) of the Bayesian network, successfully identifying implicit control loop couplings. This approach not only significantly reduces the time cost of manual network construction but also ensures the objectivity and accuracy of the network structure.

3.2 Mapping Risk Events from HAZOP to BN

Relying solely on node identifiers is insufficient to fully capture all the risks in a chemical process. To comprehensively reflect all risks and their evolution paths within the system, we match key node identifiers from the HAZOP report to potential causes and consequences of deviations, as well as the protective measures within the risk evolution paths. Using text normalization techniques, specific matching rules are established to link the risk causes, consequences, and protective measures from the HAZOP analysis table to the deviations of key variables. This enables the exploration of the risk evolution path, where risk sources cause deviations in key variables, bypass protective measures, and ultimately lead to severe consequences. This process forms a Bayesian network that incorporates risk causes, consequences, protective measures, DCS variables, and risk outcomes.

3.3 Automated Extraction of Risk Data for BN Initialization

To accurately obtain the occurrence probabilities of causes and consequences, as well as the Probability of Failure on Demand (PFD) of protection layers, the method extracts failure data from the OREDA database, SIL verification reports, and HAZOP analysis reports, constructing a comprehensive and targeted database. SIL verification reports provide failure probability analyses of protection measures for a specific unit, confirmed through expert discussions based on national standards, industry guidelines, and expert experience, thus providing targeted reliability data for protection layers. HAZOP analysis reports contain risk data on the occurrence probabilities of causes and consequences for unit-specific risk scenarios, determined through expert brainstorming sessions and analyzed according to the HAZOP procedure and expert judgment, offering targeted risk assessment data. The OREDA database, a widely used Norwegian reliability database, supplements failure probability data and risk assessment information that may not be

Table 1: Structure Example of a HAZOP table

Designator	Deviation	Cause	Protection Layer	Consequence	Probability	Severity
FI10101	High	HV10101 closed on failure	The feed will be cut off by interlock TICA10101.	Resulting in regenerator overtemperature.	3	5
TI10103	High	R0102 broke down	The settler is equipped with multiple-point temperature display.	The wax oil solidified, causing a pipeline blockage.	2	2
PIC10104	Low	P0101 broke down	P0101 one as main, one as standby.	The knockout drum was damaged due to overpressure.	3	5

captured in the SIL verification or HAZOP reports.

In the process of filling the BN with information from the database, a pre-classified Sentence-BERT (SBERT) algorithm is used to match entries, eliminating the need for manual risk probability assessment. The specific matching algorithm, as described in section 2.2, first classifies risk events into causes, protection measures, and consequences. This classification is pre-labeled according to the HAZOP analysis table during the insertion of causal events in section 3.2. The protection measures are further categorized by different equipment types, and the corresponding PFDs are matched from the database. Based on the risk event names in the Bayesian network, failure entries from the risk database are matched, allowing for the assignment of initial probabilities to risk events and completing the initialization of the Bayesian network.

3.4 Parameter Learning

After completing the initialization of the Bayesian network, parameter learning is performed using the Bayesian inference algorithm described in section 2.3. The initial probability distribution for risk event nodes is derived from data in the risk database, while a normal distribution is used as the initial probability distribution for critical node data. Actual accident occurrence data and DCS data are used as observational samples for parameter learning, allowing for the refinement of probability distributions and conditional probability tables (CPT) to better reflect real conditions.

3.5 Bayesian Inference

After the topology construction and parameter learning of the Bayesian causal network are completed, the network can be used for probabilistic inference. This network integrates risk event evidence with real-time DCS data streams to perform DRA. The inference can be performed in two modes, forward prognostic inference and backward diagnostic inference. By inputting observational data for a subset of nodes and evidence of initiating risk sources, the probabilities of different consequences can be inferred, enabling dynamic risk assessment. Conversely, given evidence of observed consequences, backward inference can be performed to trace and identify the most likely underlying causes that led to those outcomes.

4. CASE STUDY

The proposed automated Bayesian causal network construction method was applied to the fluid catalytic cracking process of a petrochemical enterprise in Zhejiang, China, to validate the feasibility of the approach.

According to the HAZOP analysis report provided by the enterprise, a total of 769 risk scenarios were identified. Among the alarm tag nodes involved in these risk events, some nodes did not have corresponding DCS data available. Consequently, 94 tag variables that were both included in the HAZOP analysis and recorded in the DCS were selected for further analysis. A Time-Series Transformer was applied to these variables for causal discovery, with the significance threshold δ set to 1.05. In total, 141 causal relationships among the variables were identified, and the resulting causal adjacency matrix is shown in Fig. 2(a). To verify the accuracy of the causal discovery algorithm, the causal relationships among these variables were manually assessed. Together with expert engineers, the production structure of the unit was analyzed to identify causal relationships based on expert knowledge, and the corresponding expert-derived causal neighborhood is illustrated in Fig. 2(b). The two matrix is generated by different ways, so the designator names are not completely same. But labels are corresponded in sequence. Treating the expert-identified causal relationships as the ground truth, a comparison with the data-driven causal matrix shows that the proposed causal discovery algorithm achieves an accuracy of 94.61%. The accuracy was calculated by the similarity

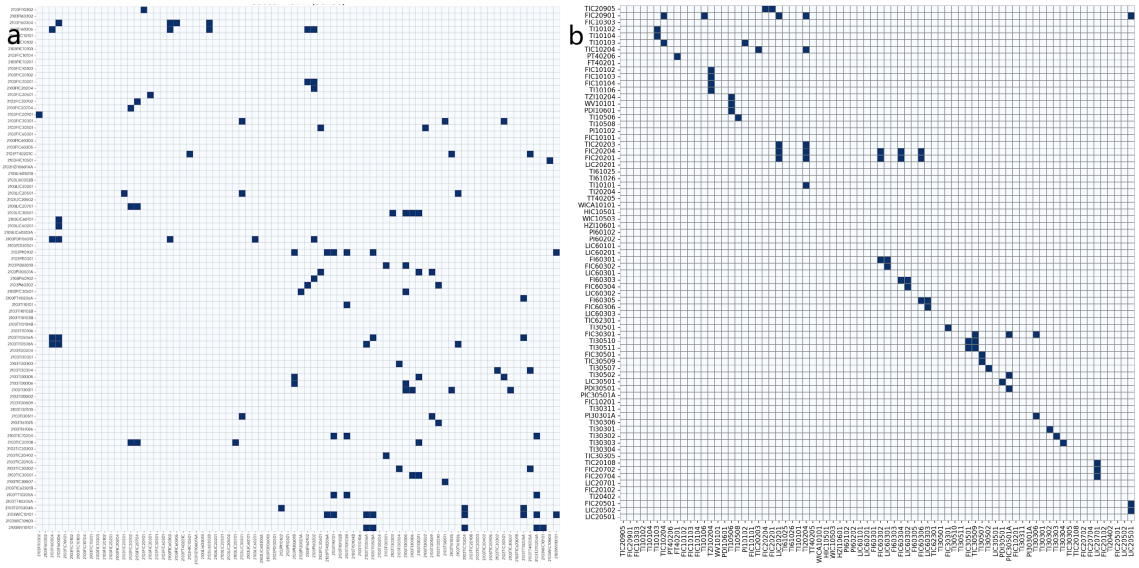

Figure 2. Causal adjacency matrix of 94 tag variables. (a) Discovered by Time-Series Transformer. (b) Discovered by expert knowledge

of two adjacent matrix.

Based on the results of causal discovery, the nodes and connections in the adjacency matrix were transformed into nodes and edges of a Bayesian network, forming its structural backbone. Using this backbone, risk events extracted from the HAZOP analysis were inserted into the Bayesian network as risk nodes, resulting in the incorporation of 260 risk scenarios. The complete Bayesian network generated is shown in Fig. 3 and consists of a total of 358 nodes, including 98 alarm tag nodes, 260 event nodes, 62 cause nodes, 61 consequence nodes, and 137 protection layer nodes, with 426 dependency relationships. The proportions of different node types in the Bayesian network are illustrated in Fig. 3.

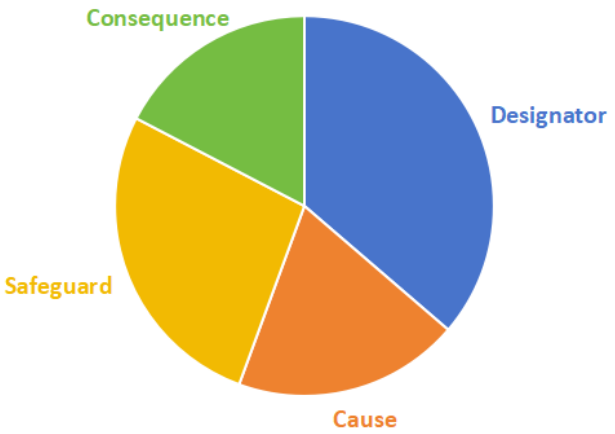

Figure 3. Nodes proportions in BN.

Based on the Bayesian network framework established using DCS time-series data and risk events, risk probabilities are matched from the risk database using a Transformer–BERT approach. The cause and consequence nodes extracted from the HAZOP analysis table can be fully and accurately associated with their corresponding probabilities. In contrast, the failure probabilities of protection layer nodes need to be matched from the risk database, with a matching accuracy of over 80%. For some protection measures, accurate and precise failure probabilities cannot be obtained from the risk database. However, this level of uncertainty is acceptable, as subsequent parameter learning is performed using actual operational data. These operational data include DCS time-series data of critical variables as well as fault records from real plant operations.

The final BN auto-constructed is shown in Fig.4, where the yellow nodes stand for hazardous causes, gray for protection layers, red for consequences, and blue for designators. This BN can be used for both forward risk inference and backward root cause tracing. Since the FCC unit considered in this case study is relatively new and no severe accident events have occurred, unplanned shutdowns are used as representative risk scenarios for analysis and validation. Taking the low–low reactor temperature interlock feed cut-off event as an example, the inference result was shown in Fig. 5. Backward risk tracing indicates that the event may have been caused by changes in feed flow rate, an increase in reactor nozzle pressure, or a failure-open condition of WICA10201. The most probable cause is lower feed flow rate, which is further caused by an increase in reactor nozzle pressure. These inferred causes are consistent with those identified in the corresponding accident analysis report.

5. CONCLUSION

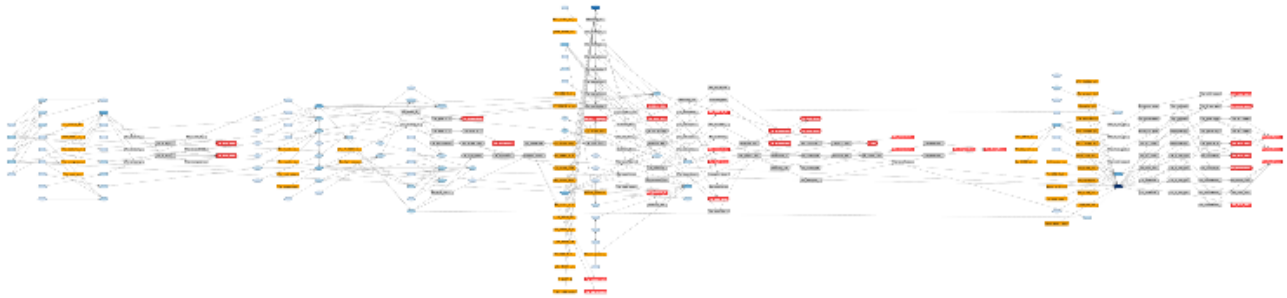

Figure 4. Auto-constructerd BN of FCC unit.

Figure 5. Risk evolution inferred by part of the BN.

This paper presents an automated framework for constructing Bayesian causal networks to enable dynamic risk assessment in complex chemical processes. By integrating time-series Transformer–based causal discovery with multi-source risk knowledge from HAZOP reports and reliability databases, the proposed approach effectively eliminates the heavy reliance on manual modeling and expert-driven network construction in conventional Bayesian network–based risk assessment.

The proposed method automatically identifies causal structures from DCS time-series data, systematically embeds risk causes, consequences, and protection layers extracted from HAZOP analyses, and assigns node probabilities through semantic matching using Sentence-BERT. Combined with Bayesian parameter learning based on real operational data, the resulting network is able to continuously adapt to process dynamics and reflect real-time risk evolution.

A case study on an industrial fluid catalytic cracking unit demonstrates that the proposed framework can accurately construct a large-scale Bayesian causal network and support both forward risk prediction and backward root cause diagnosis. The results confirm that the method significantly reduces modeling effort while providing a comprehensive and reliable representation of dynamic risk propagation in chemical plants. Overall, this work offers a practical and scalable solution for automated dynamic risk assessment and has strong potential for industrial safety management applications.

ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China under Grant 2024YFC3013600.

REFERENCES

1. Khan F, Hashemi SJ, Paltrinieri N, Amyotte P, Cozzani V, Reniers G. Dynamic risk management: a contemporary approach to process safety management. Current Opinion in Chemical Engineering 14:9-17 (2016). https://doi.org/10.1016/j.coche.2016.07.006
2. Meel A, Seider WD. Plant-specific dynamic failure assessment using bayesian theory. Chemical Engineering Science 61:7036-7056 (2006). https://doi.org/10.1016/j.ces.2006.07.007
3. Kalantarnia M, Khan F, Hawboldt K. Modelling of BP texas city refinery accident using dynamic risk assessment approach. Process Safety and Environmental Protection 88:191-199 (2010). https://doi.org/10.1016/j.psep.2010.01.004
4. Montani S, Portinale L, Bobbio A, et al. DBNet, a tool to convert dynamic fault trees into dynamic

Bayesian networks[M]//Technical Report, Dipartimento di Informatica, Università del Piemonte Orientale. 2005.
5. Kohda T, Cui W. Risk-based reconfiguration of safety monitoring system using dynamic bayesian network. Reliability Engineering & System Safety 92:1716-1723 (2007). https://doi.org/10.1016/j.ress.2006.09.012
6. Shan X, Liu K, Sun PL. Risk analysis on leakage failure of natural gas pipelines by fuzzy bayesian network with a bow-tie model. Scientific Programming 2017:1-11 (2017). https://doi.org/10.1155/2017/3639524
7. Guo X, Ji J, Khan F, Ding L, Tong Q. A novel fuzzy dynamic bayesian network for dynamic risk assessment and uncertainty propagation quantification in uncertainty environment. Safety Science 141:105285 (2021). https://doi.org/10.1016/j.ssci.2021.105285
8. Adedigba SA, Khan F, Yang M. Dynamic failure analysis of process systems using neural networks. Process Safety and Environmental Protection 111:529-543 (2017). https://doi.org/10.1016/j.psep.2017.08.005
9. Wang Q, Han J, Chen F, Zhang X, Yun C, Dou Z, Yan T, Yang G. Real-time risk prediction of chemical processes based on attention-based bi-lstm. Chinese Journal of Chemical Engineering 75:131-141 (2024). https://doi.org/10.1016/j.cjche.2024.06.026
10. Zadakbar O, Imtiaz S, Khan F. Dynamic risk assessment and fault detection using principal component analysis. Ind. Eng. Chem. Res. 52:809-816 (2012). https://doi.org/10.1021/ie202880w
11. Li S, Chen G, Men J, Li X, Zhao Y, Xu Q, Zhao J. A novel dynamic risk assessment method for hazardous chemical warehouses based on improved SVM and mathematical methodologies. Journal of Loss Prevention in the Process Industries 89:105302 (2024). https://doi.org/10.1016/j.jlp.2024.105302
12. Bi X, Wu D, Xie D, Ye H, Zhao J. Large-scale chemical process causal discovery from big data with transformer-based deep learning. Process Safety and Environmental Protection 173:163-177 (2023). https://doi.org/10.1016/j.psep.2023.03.017