CHEMICAL ENGINEERING TRANSACTIONS
VOL. 90, 2022
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Aleš Bernatík, Bruno Fabiano

ISBN 978-88-95608-88-4; **ISSN** 2283-9216
DOI: 10.3303/CET2290006

Integrating Knowledge Graph, Complex Network and Bayesian Network for Data-driven Risk Assessment

Yiping Bai*, Yuxuan Xing, Jiansong Wu

School of Emergency Management and Safety Engineering, China University of Mining & Technology, Beijing, China
Baiyiping_1995@163.com

Bayesian network is an effective method for quantitative risk assessment, but most existing studies are either heavily data-dependent or excessively expert-dependent. In this paper, knowledge graph, complex network theory and Bayesian network are integrated into a KCB model for data-driven risk assessment, especially small data situations. By applying knowledge graph with natural language processing, a causation graph could be extracted and illustrated from accident reports. Some indexes from complex network theory are introduced to identify critical nodes to simplify the huge graph. Based on the simplified network, a Bayesian network is established to quantitatively demonstrate accidents from causes to consequences. Moreover, sensitivity analysis and scenario analysis are conducted to support the decision-making of safety management. In all, the expert involvement of Bayesian network can be reduced by applying the KCB model. Besides, the KCB model can be further applied to many other areas to reach uncertainty modelling.

1. Introduction

In the field of safety science, risk can be defined as a combination of probability and severity of consequences generally (SRA, 2018). There are already many qualitative and quantitative methods for risk assessment with various applications (Aven, 2016). Based on the research of Chen et al. (2020) and the opinion of the authors, the risk assessment methods may be broadly divided into four categories: index-based methods, simulation-based methods, analytic methods and AI-based methods. With the emerging of big data and the Internet of Things (IoT), more and more data-driven risk assessment methods have been conducted in some highly digitalized domains. Hegde and Rokseth (2020) found that artificial neural network (ANN), support vector machine (SVM), decision tree (DT) and random forests (RF) are used most frequently in risk assessment. However, such machine learning approaches rely heavily on data and such a “black box” method may be difficult to explain. For the process industry, there are only small data available, which is difficult to support the traditional artificial intelligence (AI) technologies for risk assessment.

Recently, some graph-based methods like Bayesian network, Petri net, dynamic graph theory are utilized to illustrate and assess accidents of process industries (Villa et al., 2016). Due to the ability to dynamically update probabilities and can derive with limited information, the Bayesian network has been widely used. It can better model the accidents of process industries but sometimes rely on expert judgment too much, especially for some data-lacking situations. Sattari et al., (2021) applied Bayesian network and AI to conduct quantitative risk analysis and find out management priority. But such AI-based Bayesian network methods require big data to learn, which is hard to obtain. Hence, there is still a lack of a data-driven risk assessment method that can be applied for small data areas.

In this paper, knowledge graph, complex network and Bayesian network are integrated to conduct a data-driven risk assessment. First, the knowledge graph is used to illustrate the causation network of accidents. Second, some indexes of complex network theory are utilized to identify critical nodes and cut unnecessary branches. Then, a Bayesian network is established for the quantitative risk assessment. The rest of this paper is organized as follows: the methodology of this paper is elaborated in section 2; a case study based on gas pipeline accidents is demonstrated in section 3, the potential limit and future expansion of this work is discussed in section 4, and the conclusions are presented in section 5.

Paper Received: 16 November 2021; Revised: 6 March 2022; Accepted: 27 April 2022
Please cite this article as: Bai Y., Xing Y., Wu J., 2022, Integrating knowledge graph, complex network and Bayesian network for data-driven risk assessment, Chemical Engineering Transactions, 90, 31-36 DOI:10.3303/CET2290006

2. Methodology

The knowledge graph, complex network, Bayesian network and the integrated KCB (Knowledge graph - Complex network - Bayesian network) model are introduced as follows.

2.1 Knowledge graph

Knowledge graph is born to illustrate the relationship between each entity and firstly used in search engines. A knowledge graph is a multi-node (entity) graph connected by directed edges (relations) with properties and values (Yu et al., 2021). Generally, the establishment of a knowledge graph consists of three parts: knowledge extraction, knowledge fusion and knowledge storage. Recently, rule-based matching, machine learning are commonly used methods for building a knowledge graph.

2.2 Complex network

Complex network theory is developed based on graph theory and statistical physics. In complex network theory, every complex system can be abstracted as a network. The nodes in the network can be regarded as the elements in the system. If the various elements in the system are regarded as nodes and the relations between each element as connections, then the system constitutes a network. There are many structural indicators in complex network theory which can quantify the importance of nodes from the node itself or the network relationship level, such as degree centrality, eigenvector centrality, and clustering coefficient (Kim and Perez, 2015).

2.3 Bayesian network

Bayesian network is also called Bayesian belief network. It is a directed acyclic graph based on probabilistic reasoning and prediction to express and analyze uncertain events (Pearl, 1988). It consists of a variety of different nodes and directed edges. Each node represents a variable (which can be an observable variable, a hidden variable, an unknown parameter, etc.). There are two kinds of nodes, parent nodes and child nodes. And the directed edges represent the dependency relationship between the nodes (from a parent node to a child node). The Bayesian network is a kind of probability model to illustrate causation network and is widely used in almost every area.

2.4 Integrated KCB model

The integrated KCB (Knowledge graph - Complex network - Bayesian network) model contains three methods and five steps. First, knowledge graph is applied with natural language processing (NLP) to illustrate causation network from accident reports. Second, all sub-graph of the established knowledge graph is fused to a comprehensive network. Third, several indexes of the complex network are used to quantify importance of each node and find out critical ones to simplify the huge network. Fourth, a Bayesian network based on the simplified network is established. Finally, quantitative risk analysis is conducted based on sensitivity analysis and scenario analysis of the proposed Bayesian network.

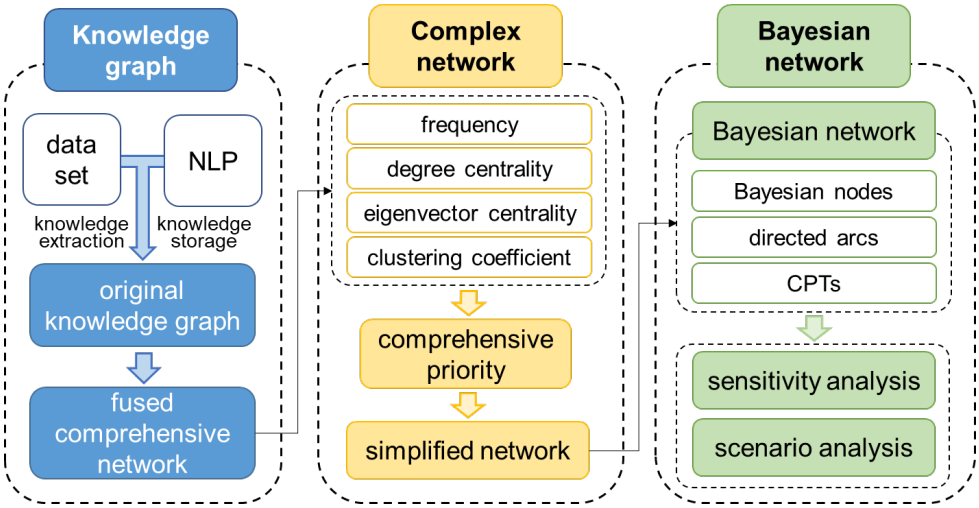

Figure 1: Flow chart of the integrated KCB model

3. Case study

In this paper, the buried gas pipeline accidents from 2018-2021 in China are collected to conduct a case study of the proposed KCB model. The establishment of knowledge graph, simplification based on complex network, the establishment of Bayesian network, sensitivity analysis and scenario analysis of the proposed Bayesian network are elaborated in sequence as follows.

3.1 Establishment of knowledge graph

First, the accident reports of directly buried gas pipelines are collected from the Internet to build a data set, which contains 124 gas pipeline accidents from 2018 to 2021 in China. Then, by applying LTP (LTP, 2014), a widely used natural language processing tool to conduct word segmentation, speech tagging, dependency parsing and role labelling, the original knowledge graph is established as Figure 2a shows. However, the original knowledge graph consists of many sub-graphs, which cannot be transferred to Bayesian network. Hence, the original graph is then fused to form a comprehensive graph as Figure 2b shows. The fusion is based on the primary pattern of gas pipeline accidents “causes (in green) -gas leakage (in blue) -accidents (in purple) -losses (in red)”. Not only repeated nodes of the knowledge graph are merged, but the relationship between cause nodes in different event chains is also determined by literature research and expert consultation.

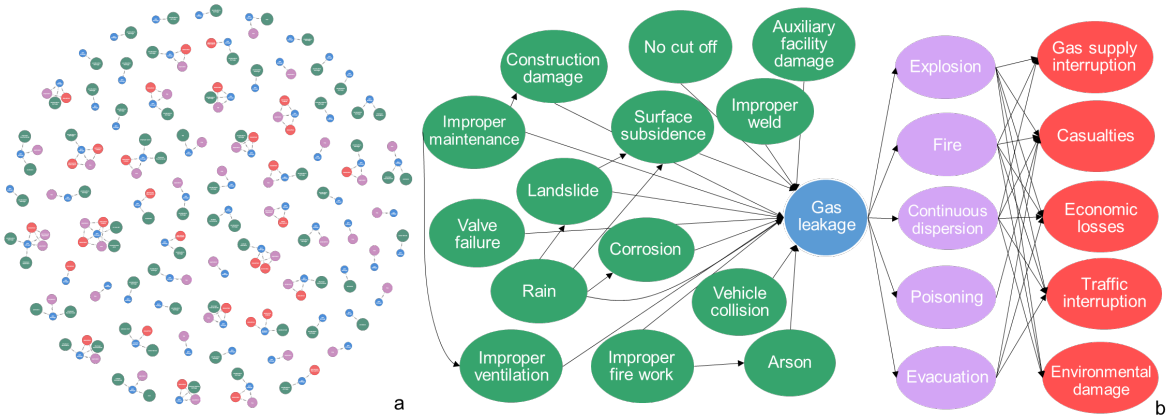

Figure 2: Knowledge graph of gas pipeline accidents before (2a) and after (2b) fusion

3.2 Optimization and simplification based on complex network

For clearer illustrating the topology of gas pipeline accidents and better establishing Bayesian network, it is necessary to simplify (i.e., cut branches) the proposed knowledge graph. Therefore, the complex network theory is introduced to optimize and simplify the graph by calculating several importance indexes. The frequency ratio, degree centrality, eigenvector centrality and clustering coefficient are calculated to form a comprehensive priority ratio. The simplification of the graph structure varies by different types of nodes. For the accident nodes and loss nodes, there is no interconnection within the various nodes, so the accident nodes and loss nodes are grouped into one category for simplification in this paper and the calculation of the clustering coefficient is omitted. Because of the higher the structural importance of the node, the probability of its occurrence is not necessarily higher. Therefore, in addition to the calculation of the key indicators of the complex network, the occurrence frequency of each node is also normalized to obtain the frequency ratio. The weights of key indicators are determined through the Analytic Hierarchy Process (AHP). For the cause node, the weights of the indicators are W_{FR}:W_{DC}:W_{EC}:W_{CC}=0.65:0.1:0.1:0.15. For accident nodes and loss nodes, the weights of the indicators are W_{FR}:W_{EC}:W_{CC}=0.7:0.15:0.15. Therefore, the comprehensive priority score of each node is obtained by weighting (Table 1). By ranking the comprehensive priority score of nodes in each type, the important nodes can be identified to support further assessment. With different purposes and objects, a different number of nodes could be filtered. In this case, the top four cause nodes, top three accident nodes and top three loss nodes are selected as critical nodes to form the simplified network.

Table 1: Importance indexes of nodes in the fused knowledge graph

Node	Frequency ratio	Degree centrality	Eigenvector centrality	Clustering coefficient	Comprehensive priority score
Construction Damage	0.569	0.154	0.083	0.667	0.494
Improper Maintenance	0.108	0.308	0.125	0.500	0.188
Corrosion	0.077	0.154	0.083	0.667	0.174
Valve Failure	0.031	0.154	0.083	0.667	0.144
Surface Subsidence	0.031	0.231	0.146	0.833	0.183
Improper Ventilation	0.031	0.154	0.083	0.667	0.144
Auxiliary Facilities Damage	0.031	0.077	0.000	0.000	0.028
Arson	0.031	0.154	0.042	0.667	0.140
Improper Fire Work	0.015	0.154	0.042	0.667	0.130
Landslide	0.015	0.231	0.146	0.833	0.173
Improper Weld	0.015	0.077	0.000	0.000	0.018
Vehicle Collision	0.015	0.077	0.000	0.000	0.018
Rain	0.015	0.308	0.167	0.700	0.162
No Cut Off	0.015	0.077	0.000	0.000	0.018
Explosion	0.542	0.556	0.253		0.501
Fire	0.375	0.556	0.253		0.384
Continuous Dispersion	0.042	0.556	0.253		0.151
Poisoning	0.021	0.111	0.067		0.041
Evacuation	0.021	0.333	0.173		0.091
Gas Supply Interruption	0.417	0.333	0.176		0.368
Casualties	0.250	0.556	0.224		0.292
Economic Losses	0.167	0.444	0.212		0.215
Traffic Interruption	0.083	0.444	0.212		0.157
Environmental Damage	0.083	0.333	0.176		0.135

3.3 Establishment of Bayesian network

Based on the structure of the simplified network in the previous section, the topology structure of Bayesian network is established. Then, the prior probabilities of parent nodes and conditional probabilities of child nodes are determined by statistics and expertise with the Delphi method, which is a multi-feedback expert communication method to make sure the expert opinions in consistence. As a result, the Bayesian network with the ability to dynamically update posterior probabilities is established as shown in Figure 3. With the proposed Bayesian network, the sensitivity analysis and scenarios analysis are conducted in the following section.

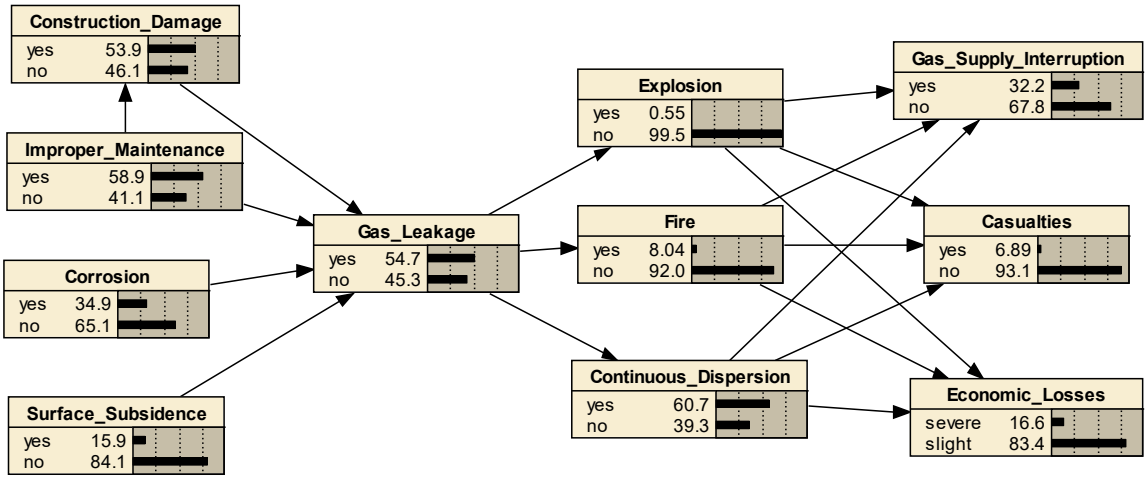

Figure 3: Bayesian network based on the simplified network

3.4 Sensitivity analysis of the Bayesian network

The sensitivity analysis of Bayesian network can quantify the influence of each parent node that affect a certain child node. For the Bayesian network of gas pipeline accidents, it is essential to find out which cause is most likely to result in gas leakage. By calculating the sensitivity of the three cause nodes to the gas leakage node, the results of Table 2 shows that “Construction damage” and “Improper maintenance” can dramatically impact the occurrence of gas leakage, the sensitive proportion of them are 17.7 % and 17.6 %, respectively. Similarly, the main accident that affects serious casualties is “Fire”, whose sensitive proportion is 35.6 %. And both gas supply interruption and serious economic loss are mostly affected by “Continuous dispersion” with sensitivities higher than 28 % and 13 %, respectively (Table 2). Hence, these nodes with higher sensitivities need to be handled with more effort.

Table 2: Sensitivity analysis results of Bayesian network

target nodes	Sensitive node	Sensitive proportion
Gas leakage	Construction damage	17.7 %
Interruption	Improper maintenance	17.6 %
	Corrosion	4.81 %
	Surface subsidence	2.03 %
Casualties	Fire	35.6 %
	Continuous dispersion	12.6 %
	Explosion	2.75 %
Gas supply	Continuous dispersion	28.5 %
	Fire	9.25 %
	Explosion	0.712 %
Economic losses	Continuous dispersion	13.3 %
	Fire	3.01 %
	Explosion	0.281 %

3.5 Scenario analysis of the Bayesian network

In this section, evidence (one or more nodes setting in a certain state) is given to calculate the posterior probabilities of the proposed Bayesian network. Based on the previous sensitivity analysis, “Construction damage”, “Fire” and “Continuous dispersion” are the key nodes in this Bayesian network. Combing the states of the three key nodes, seven potential scenarios of gas pipeline accidents are designed and their results are shown in Table 3. Meanwhile, the Bayesian network with updated posterior probabilities in scenario 2 is shown in Figure 4. Through the scenario analysis, not only the consequence probability of a particular scenario can be calculated, the impact of different nodes can be quantified by comparing different scenarios. The results show that preventing fire when the construction damage node occurs can effectively reduce the accident loss (supply interruption from 87.5 % to 47 % and casualties from 57.9 % to 4.35 %). Meanwhile, although the continuous dispersion cannot directly cause great casualties, the probability of triggered gas supply interruption is still higher than 47 %.

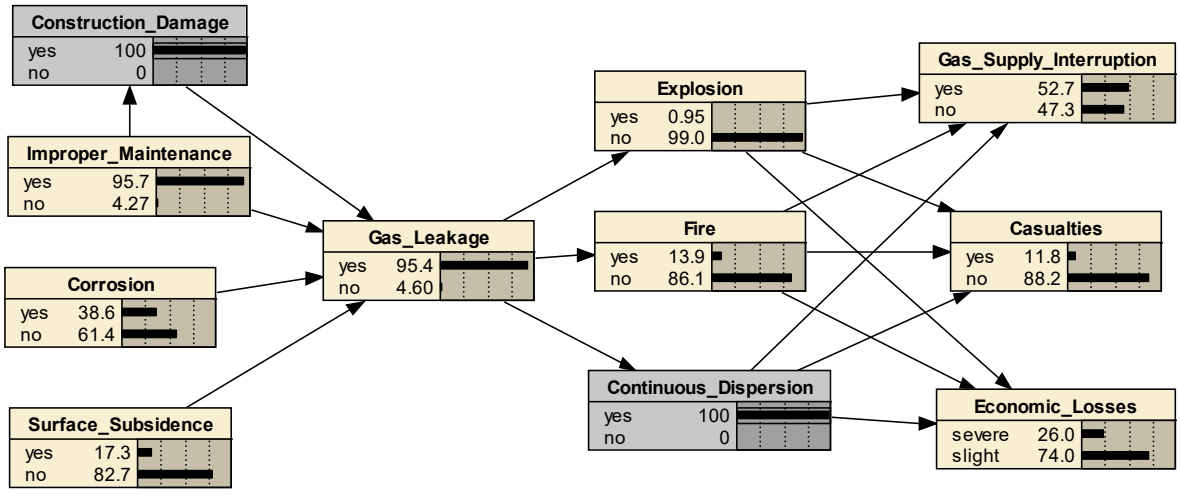

Figure 4: Bayesian network in scenario 2

Table 3: The states and results of scenario analysis

Scenarios	Construction damage	Continuous dispersion	Fire	Gas supply interruption	Casualties	Economic losses
1	yes	yes	yes	87.5 %	57.9 %	39.8 %
2	yes	yes	no	47.0 %	4.35 %	23.8 %
3	yes	no	yes	74.8 %	37.9 %	35.6 %
4	yes	no	no	1.06 %	0.043 %	2.13 %
5	yes	yes	unknow	52.7 %	11.8 %	26.0 %
6	yes	unknow	yes	87.2 %	57.4 %	39.7 %

4. Discussions

Due to the limited space, the Bayesian network is small, the sensitivity analysis and the scenario analysis is relatively simple, and the conditional probability tables still rely on expert judgment. However, many works could be further analyzed. With more scenarios designed and analyzed in the Bayesian network, the results can effectively and quantitatively support the design of the emergency plan. If the sensors of gas pipelines could be linked with the proposed model, the Bayesian network can dynamically update and support real-time emergency decision-making. By adding safety barrier nodes and comparing results before and after improvement, the optimization strategy could be put forward.

Through the gas pipeline accident case study, the proposed KCB model is proved useful for risk assessment with small data. Moreover, this model can be applied for both big data and small data, as long as textual data with causation relationships. Besides, the KCB model can be applied in almost every area, process industries, construction industries, economics, medicine, etc. Also, the three methods in the KCB model can be flexibly changed to better focus on the analyzed issue.

5. Conclusions

In this paper, knowledge graph, complex network theory and Bayesian network are integrated to form a KCB model for data-driven risk assessment. With the proposed model, knowledge (factors and relationships) could be extracted and illustrated from small data. And the causation topology can be quantitatively demonstrated with Bayesian network. The subjectivity of the establishment of Bayesian network can be relatively avoided by applying the KCB model. And the KCB can be easily adjusted to suit many other issues.

A case study for buried gas pipeline accidents in China is analyzed based on the proposed KCB model. A knowledge graph with 241 nodes and 155 edges is put forward and further transformed to an 11-node Bayesian network by complex network theory. Results of sensitivity analysis and scenario analysis of the Bayesian network provide quantitative reference to safety management: controlling fire after construction damage occurs can reduce the probability of supply interruption from 87.5 % to 47 % and casualties from 57.9 % to 4.35 %. Moreover, the results indicate the proposed model can suit not only big data but also small data, which is especially practical for some traditional industries.

References

Aven T., 2016, Risk assessment and risk management: Review of recent advances on their foundation, European Journal of Operational Research, 253, 1-13.

Chen C., Reniers G., Khakzad N., 2020, A thorough classification and discussion of approaches for modeling and managing domino effects in the process industries, Safety Science, 125, 104618.

Hegde J., Rokseth B., 2020, Applications of machine learning methods for engineering risk assessment – A review, 122, 104492.

Kim J. and Perez C., 2015, Co-Authorship Network Analysis in Industrial Ecology Research Community. Journal of Industrial Ecology, 19, 222-235.

LTP, 2014, Language Technology Platform. <http://www.ltp-cloud.com/intro> accessed 26.11.2021.

Pearl J., 1988, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, US.

Sattari F., Macciotta R., Kurian D., Lefsrud L., 2021, Application of Bayesian network and artificial intelligence to reduce accident/incident rates in oil & gas companies, Safety Science, 133, 104981.

SRA, 2018, Society for risk analysis glossary. <https://www.sra.org/wp-content/uploads/2020/04/SRA-Glossary-FINAL.pdf> accessed 14.11.2021.

Villa V., Paltrinieri N., Khan F., Cozzani V., 2016, Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry, Safety Science, 89, 77-93.

Yu C., Wang F., Liu Y.H., An L., 2021, Research on knowledge graph alignment model based on deep learning, Expert Systems with Applications, 186, 115768.