TU Delft

Delft University of Technology

Bayesian network model to distinguish between intentional attacks and accidental technical failures

a case study of floodgates

Chockalingam, Sabarathinam; Pieters, Wolter; Teixeira, André; van Gelder, Pieter

DOI
10.1186/s42400-021-00086-6

Publication date
2021

Document Version
Final published version

Published in
Cybersecurity

Citation (APA)
Chockalingam, S., Pieters, W., Teixeira, A., & van Gelder, P. (2021). Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates. *Cybersecurity*, *4*(1), Article 29. https://doi.org/10.1186/s42400-021-00086-6

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Chockalingam et al. Cybersecurity (2021) 4:29
https://doi.org/10.1186/s42400-021-00086-6

Cybersecurity

RESEARCH Open Access

Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates

Sabarathinam Chockalingam[1,2*], Wolter Pieters[1,3], André Teixeira[4] and Pieter van Gelder[1]

Abstract

Water management infrastructures such as floodgates are critical and increasingly operated by Industrial Control Systems (ICS). These systems are becoming more connected to the internet, either directly or through the corporate networks. This makes them vulnerable to cyber-attacks. Abnormal behaviour in floodgates operated by ICS could be caused by both (intentional) attacks and (accidental) technical failures. When operators notice abnormal behaviour, they should be able to distinguish between those two causes to take appropriate measures, because for example replacing a sensor in case of intentional incorrect sensor measurements would be ineffective and would not block corresponding the attack vector. In the previous work, we developed the attack-failure distinguisher framework for constructing Bayesian Network (BN) models to enable operators to distinguish between those two causes, including the knowledge elicitation method to construct the directed acyclic graph and conditional probability tables of BN models. As a full case study of the attack-failure distinguisher framework, this paper presents a BN model constructed to distinguish between attacks and technical failures for the problem of incorrect sensor measurements in floodgates, addressing the problem of floodgate operators. We utilised experts who associate themselves with the safety and/or security community to construct the BN model and validate the qualitative part of constructed BN model. The constructed BN model is usable in water management infrastructures to distinguish between intentional attacks and accidental technical failures in case of incorrect sensor measurements. This could help to decide on appropriate response strategies and avoid further complications in case of incorrect sensor measurements.

Keywords: Bayesian network, DeMorgan model, Intentional attack, Probability elicitation, Safety, Security, Technical failure, Water management

Introduction

Water management is one of the critical infrastructures in countries like the Netherlands (Castellon and Frinking 2015). The proper functioning of water management infrastructures is vital for economic growth and societal wellbeing. The unexpected closure of floodgates could lead to severe economic damage, for instance, by delaying cargo ships. Over the years, water management infrastructures have become dependent on Industrial Control Systems (ICSs) to ensure efficient operations of such infrastructures (Nogueira and Walraven 2018).

ICSs were originally designed for isolated environments (Effendi and Davis 2015). Such systems were mainly susceptible to technical failures. The blackout in the Canadian province of Ontario and the North-eastern

* Correspondence: Sabarathinam.Chockalingam@ife.no
[1]Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands
[2]Department of Risk, Safety and Security, Institute for Energy Technology, Halden, Norway
Full list of author information is available at the end of the article

Springer Open

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

and Mid-western United States is a typical example of a technical failure in which the absence of alarm due to a software bug in the alarm system left operators unaware of the need to redistribute power (Zhivich and Cunningham 2009). However, modern ICSs no longer operate in isolation, but use other networks to facilitate and improve business processes (Knowles et al. 2015). This increased connectivity makes ICSs more vulnerable to cyber-attacks apart from technical failures. A cyber-attack on a German steel mill is a typical example in which adversaries made use of corporate network to enter the ICS network (RISI 2014). As an initial step, the adversaries used both the targeted email and social engineering techniques to acquire credentials for the corporate network. Once they acquired credentials for the corporate network, they worked their way into the plant's control system network and caused damage to the blast furnace.

It is essential to distinguish between attacks and technical failures that would lead to abnormal behaviour in the components of ICSs and take suitable measures. In most cases, the initiation of response strategy presumably aimed at technical failures would be ineffective in the event of a targeted attack and may lead to further complications. For instance, replacing a water level sensor that is sending incorrect measurement data with a new water level sensor would be a suitable response strategy to technical failure of a water level sensor. However, this may not be an appropriate response strategy to an attack on the water level sensor as it would not block the corresponding attack vector. Furthermore, the initiation of inappropriate response strategies would delay the recovery of the system from adversaries and might lead to harmful consequences. Noticeably, there is a lack of decision support to distinguish between attacks and technical failures.

Bayesian Networks (BNs) have the capacity to tackle this challenge especially based on their real-world applications in medical diagnosis and fault diagnosis (Nakatsu 2009). BNs belong to the family of probabilistic graphical models, consisting of a qualitative and a quantitative part (Darwiche 2008). The qualitative part is a Directed Acyclic Graph (DAG) of nodes and edges. Each node represents a random variable, while the edges between the nodes represent the conditional dependencies among the random variables. The quantitative part takes the form of a priori marginal and conditional probabilities so as to quantify the dependencies between connected nodes.

In order to address the above-mentioned research gap, we developed the attack-failure distinguisher framework in our previous work to help construct BN models for distinguishing attacks and technical failures (Chockalingam et al. 2019; Chockalingam et al. 2020). Furthermore, we extended and combined fishbone diagrams within our framework for knowledge elicitation to construct the qualitative part of such BN models. Finally, we integrated DeMorgan models and probability scales with numerical and verbal anchors within our framework for knowledge elicitation to construct the quantitative part of such BN models. The present study aims to construct a BN model based on the developed framework to distinguish between attacks and technical failures for an observable problem in floodgates, providing a full case study of the framework as well as addressing the problem of floodgate operators. This paper addresses the research question: *"How could we develop Bayesian Network (BN) models for distinguishing attacks and technical failures in floodgates?"*. The research objectives are:

RO1. To develop a BN model for distinguishing attacks and technical failures in floodgates involving domain experts using the attack-failure distinguisher framework.
RO2. To demonstrate the suitability of a BN model developed with the attack-failure distinguisher framework in floodgates.

RO1 focuses primarily on the use of the expert elicitation process proposed in the attack-failure distinguisher framework to develop a BN model for distinguishing attacks and technical failures in floodgates. Even though the available system information during the elicitation process is limited, this would not have an impact on providing a full case study of the framework. RO2 focuses mainly on demonstrating when and how a BN model developed with the framework would be useful in practice, and not on assessing the validity of the specific BN model, due to the lack of real water management infrastructure and testbed for evaluation.

At the start of this research, we investigated the availability of data corresponding to cyber-attacks and technical failures from real-world systems in the water management sector. This data would help to construct DAGs and populate Conditional Probability Tables (CPTs). However, there is a lack of data regarding cyber-attacks from real-world systems as experts in safety and/or security of ICS in the water management sector in the Netherlands claim that there are no/limited cyber-attacks on their infrastructures. These experts are associated with the organisation responsible for the construction and maintenance of flood protection and prevention in the Netherlands and their suppliers. Moreover, data corresponding to limited cyber-attacks that happened is not shareable due to the sensitivity of data.

On the other hand, technical failures occur in their infrastructures which are documented as technical failure reports. However, they are also not shareable due to the

sensitivity of data. Therefore, we relied on expert knowledge which is one of the predominant data sources utilised to construct DAGs and populate CPTs especially in domains where there is a limited availability of data like cyber security (Chockalingam et al. 2017). Furthermore, expert knowledge is substantive information on a specific domain based on the system knowledge that is not commonly known by others (Martin et al. 2012). Finally, it is also prevalent to use expert knowledge as the data source which is one of the well-established and successful alternate data source to data from real-world systems in modelling cyber security (Holm et al. 2013; Husák et al. 2018). Specifically, we utilised experts who associate themselves with safety and/or security community as it is appropriate for our application which deals with distinguishing attacks and technical failures. In our context, we associate the security community as dealing with attacks. On the other hand, we associate the safety community as dealing with technical failures.

The main contributions of this paper are as follows:

(i) we provide a full case study of the attack-failure distinguisher framework on how to construct a BN model for an observable problem in floodgates using expert knowledge.
(ii) we develop decision support that help operators to distinguish between attacks and technical failures for the problem of incorrect sensor measurements in floodgates in the Netherlands.
(iii) we demonstrate the suitability of the constructed BN model in the water management sector by showing when and how this could be used in practice.

The remainder of this paper is structured as follows. In Section 2, we illustrate the different layers and the components of an ICS. In Section 3, we describe our existing framework that would help to construct BN models for distinguishing attacks and technical failures in addition to the systematic methods for knowledge elicitation to construct the BN models. Section 4 demonstrates the constructed BN model followed by discussions in Section 5. Section 6 highlights the related work.

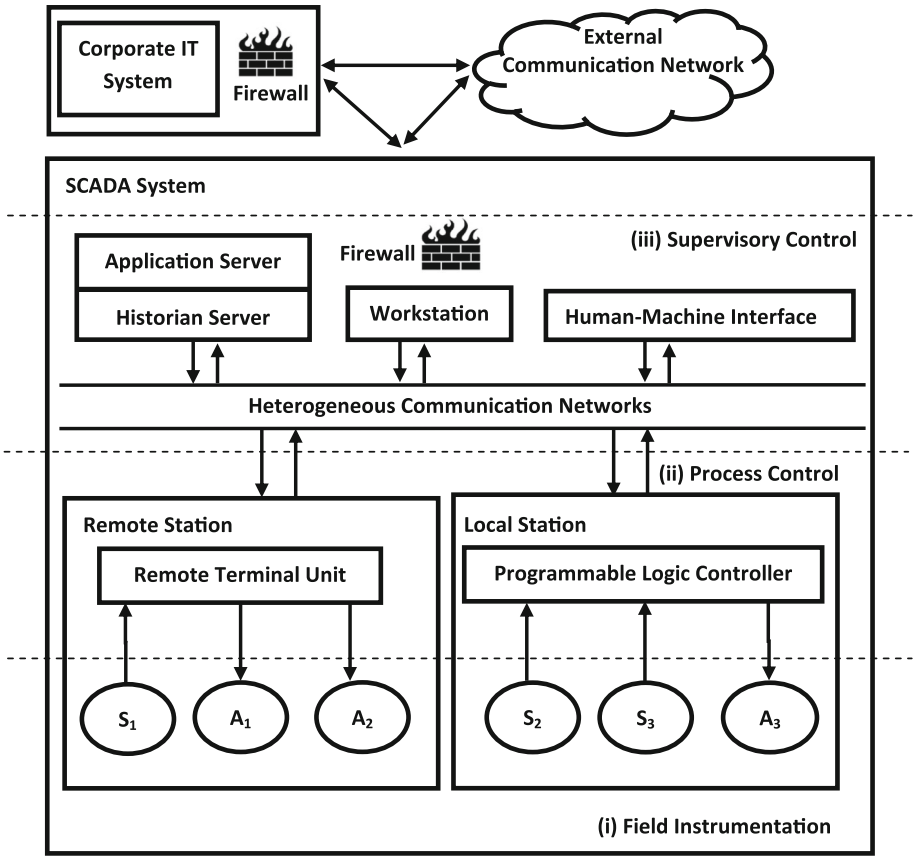

Fig. 1 Typical ICS Architecture and Layers

Section 7 presents the conclusions and future work directions.

ICS architecture

In this section, we illustrate the three different layers and major components in each layer of an ICS.

Domain knowledge on ICSs is the starting point for the application of our proposed approach. A typical ICS consists of three layers: (i) Field instrumentation, (ii) Process control, and (iii) Supervisory control, bound together by network infrastructure, as shown in Fig. 1.

The field instrumentation layer consists of sensors (S_i) and actuators (A_i), while the process control layer consists of Programmable Logic Controllers (PLCs)/Remote Terminal Units (RTUs). Typically, PLCs have wired communication capabilities whereas RTUs have wired or wireless communication capabilities. The PLC/RTU receives measurement data from sensors, and controls the physical systems through actuators (Skopik and Smith 2015). The supervisory control layer consists of historian databases, software application servers, the Human-Machine Interface (HMI), and the workstation. The historian databases and software application servers enable the efficient operation of the ICS. The low-level components are configured and monitored with the help of the workstation and the HMI, respectively (Skopik and Smith 2015).

Framework for distinguishing attacks and technical failures

This section describes the attack-failure distinguisher framework proposed in our previous work to construct BN models for distinguishing attacks and technical failures (Chockalingam et al. 2019).

The framework consists of three layers as shown in Fig. 2. The middle layer consists of a problem variable which is the major cause for an abnormal behaviour in a component of the ICS (observable problem). The states of the problem variable are the major causes of the observable problem (intentional attack and accidental technical failure). The upper layer consists of factors contributing to the major causes of the problem. The lower layer consists of observations (or test results) which is defined as any information useful for determining the major cause of the problem based on the outcome of tests conducted once the problem is observed by a floodgate operator.

The BN models would be incomplete without the quantitative part (CPTs for each variable). However, probability elicitation is a challenging task in building BNs, especially when it relies heavily on expert knowledge (Zhang and Thai 2016). The extensive workload for experts in probability elicitation could affect the reliability of elicited probabilities. Therefore, the framework which we proposed in our previous work also includes DeMorgan models that reduces the number of conditional probabilities to elicit from domain experts in

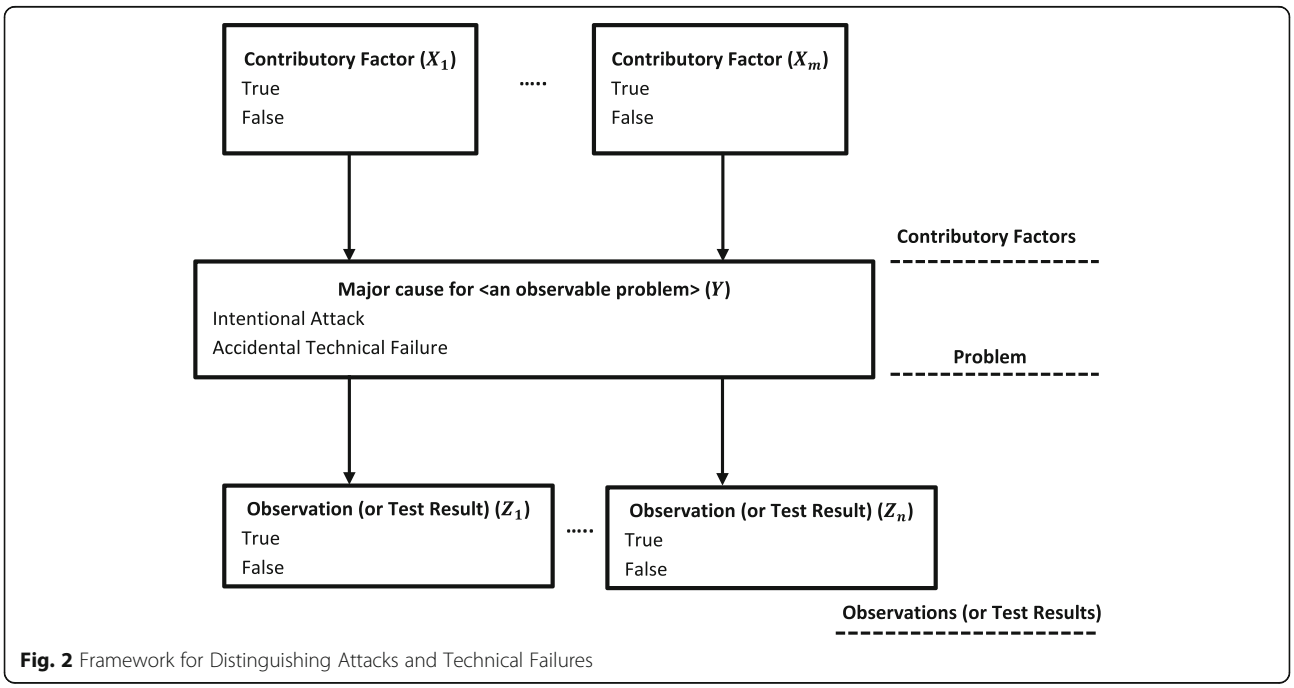

Fig. 2 Framework for Distinguishing Attacks and Technical Failures

constructing the quantitative part of BN models, especially this technique reduces the number of parameters that need to be elicited from exponential to linear in the number of parents to define a full CPT for the child variable (Chockalingam et al. 2019; Chockalingam et al. 2020). We adopted DeMorgan models because it is the most suitable technique for our purpose (Chockalingam et al. 2020). Furthermore, we integrated probability scales with numerical and verbal anchors with DeMorgan models to facilitate individual probability entry by providing visual aids to help experts answer in terms of probabilities (Chockalingam et al. 2020).

The DeMorgan model is applicable when there are several parents and a common child. The DeMorgan model inherently assumes binary variables. In our application, the DeMorgan model could be used to elicit conditional probabilities for the problem variable as they have several contributory factors (parents). On the other hand, the CPTs of the contributory factors and observations (or test results) could be elicited directly from experts as they are straightforward when they do not have several parents. The DeMorgan model assumes that one of the two states of each variable is always the distinguished state as shown in Fig. 3. Usually such state of the child variable depends on the modelled domain (Zagorecki 2010). This is a typical state of the corresponding child variable (Kraaijeveld 2005). In our application, the distinguished state of the problem variable ("Major cause for <an observable problem>") is chosen as "accidental technical failure" as this is the a priori expected major cause, based on the higher frequency of technical failures compared to the attacks (Chockalingam et al. 2019; Chockalingam et al. 2020). The distinguished state of a parent variable is relative to the type of causal interaction with the child variable (Maaskant and Druzdzel 2008). The same parent variable can have different distinguished states in different interactions that it participates in with the different child variables.

There are four different types of causal interactions between an individual parent (X) and a child (Y) in the DeMorgan model: (i) cause, (ii) barrier, (iii) inhibitor, and (iv) requirement.

(i) Cause: X is a causal factor and has a positive influence on Y. In this type of causal interaction between an individual parent (X) and a child (Y), the distinguished state of the corresponding parent variable is "False" (Maaskant and Druzdzel 2008). Consequently, when the parent variable is "False", it is certain not to trigger a change from the typical state of the child variable. When the parent variable is "True", it will trigger a change from the typical state of the child variable, with a certain probability (v_X).

(ii) Barrier: This is a negated counterpart of cause, i.e., X' is a causal factor and has a positive influence on Y. In this type of causal interaction between an individual parent (X) and a child (Y), the distinguished state of the corresponding parent variable is "True" (Maaskant and Druzdzel 2008). Accordingly, when the parent variable is "True", it is certain not to trigger a change from the typical state of the child variable. When the parent variable is "False", it will trigger a change from the typical state of the child variable, with a certain probability (v_X).

(iii) Inhibitor: X inhibits Y. In this type of causal interaction between an individual parent (X) and

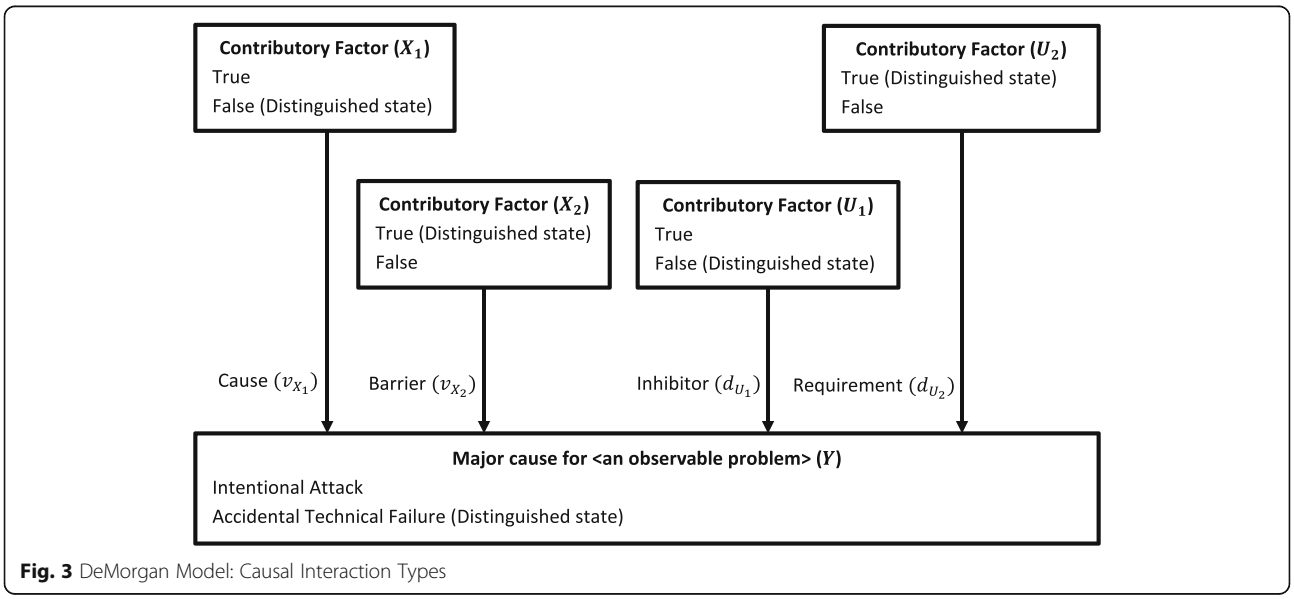

Fig. 3 DeMorgan Model: Causal Interaction Types

a child (Y), the distinguished state of the corresponding parent variable is "False" (Maaskant and Druzdzel 2008). As a result, when the parent variable is "False", it is certain not to prevent a change from the typical state of the child variable. When the parent variable is "True", it will prevent a change from the typical state of the child variable, with a certain probability (d_X).

(iv) Requirement: The relationship between an inhibitor and requirement is similar to the relationship between a cause and barrier. X' inhibits Y. In this type of causal interaction between an individual parent (X) and a child (Y), the distinguished state of the corresponding parent variable is "True" (Maaskant and Druzdzel 2008). Hence, when the parent is "True", it is certain not to prevent a change from the typical state of the child variable. When the parent variable is "False", it will prevent a change from the typical state of the child variable, with a certain probability (d_X).

The DeMorgan model is an extension and a combination of the noisy-OR and noisy-AND model which supports modelling the above-mentioned types of causal interactions (Maaskant and Druzdzel 2008). The property of accountability in the noisy-OR model is applicable to the DeMorgan model with a slight modification as it also exploits causal independence: In case all the modelled parents of the child are in their distinguished state, the property of accountability requires that the child be presumed their distinguished state. However, in many cases, this is not a realistic assumption as it is difficult to capture all the possible parents of the child (Fallet-Fidry et al. 2012). Specifically, this is not realistic in our application as it is difficult to capture all the possible contributory factors of an observable problem due to "intentional attack". In the DeMorgan model, the leak parameter (ν_{X_L}) deals with the possible parents of the child that are not previously known and explicitly modelled.

In general, the size of the CPT of a binary variable with n binary parents is 2^{n+1}. However, only $n+1$ parameters are sufficient to completely define CPT using the DeMorgan model as it exploits causal independence. In the example shown in Fig. 3, only five parameters are sufficient to completely define the CPT of child variable (Y) using the DeMorgan model instead of 64 entries. We could find the values for required parameters from the experts to completely define CPT using the DeMorgan model based on appropriate question for each type of causal interaction shown in Table 1.

Once we determine the required parameters based on appropriate elicitation questions, we can completely define the CPT of the child variable using (1):

$$P(y|X,U) = \left(1-(1-\nu_{X_L}) \prod_{X_i \in +X} (1-\nu_{X_i})\right) \prod_{U_i \in +U} (1-d_{U_i}) \tag{1}$$

In the Eq. (1), Y represents the effect variable which has values y for the effect being in the non-distinguished state ("Intentional attack") and y' for the effect being in the distinguished state ("Accidental technical failure"). X denotes the set of parents which interact with the effect variable as promoting influences, U denotes the set of parents which interact with the effect variable as inhibiting influences, $+X$ denotes the subset of X that contains all parents that are in their non-distinguished states, $+U$ denotes the subset of U that contains all parents that are in their non-distinguished states. ν_{X_L} denotes the leak parameter which expresses the probability of y ("Intentional attack") given all parents are in their distinguished states, ν_{X_i} denotes the probability of y ("Intentional attack") given that the parent X_i is not in its distinguished state and all other parents are in their distinguished states, d_{U_i} denotes the probability of y' ("Accidental technical failure") given that the parent U_i is not in its distinguished state and all other parents are in their distinguished states.

Applying BNs for distinguishing attacks and technical failures

This section describes how we constructed the BN model for distinguishing attacks and technical failures in

Table 1 Causal Interactions and their Corresponding Elicitation Questions in the DeMorgan Model

Type of Causal Interaction	Elicitation Question
Leak	*"What is the probability that the child is in their non-distinguished state given that the parents are in their distinguished states?"*
Cause, Barrier Note: There is a difference between the non-distinguished state of a cause and barrier.	*"What is the probability that the child is in their non-distinguished state given that all the parents are in their distinguished states, except X_i and no other unmodelled causal factors are present?".*
Inhibitor, Requirement Note: There is a difference between the non-distinguished state of an inhibitor and requirement.	*"What is the probability that the child is in their distinguished state given that the parents are in their distinguished states, except U_i and no other unmodelled causal factors are present?".*

floodgates. We considered the observable problem for this application as "Sensor sends incorrect water level measurements" because it could lead to serious consequences in the case of floodgate. In case the floodgate closes when it should not, based on the incorrect water level measurements sent by the sensor, it would lead to severe economic damage, for instance, by delaying cargo ships. On the other hand, in case the floodgate opens when it should not, due to incorrect water level measurements sent by the sensor, it would lead to flooding.

Construction of qualitative BN model for distinguishing attacks and technical failures in floodgates

We have utilised a multimethodology approach for data collection. Multimethodology refers to using more than one method of data collection in a research study (Brewer and Hunter 1989), providing more comprehensive data. In our study, we utilised a focus group workshop and a questionnaire to gather data for constructing the qualitative BN model. Firstly, we conducted a focus group workshop with five participants who have experience working with safety and/or security of water management infrastructures operated by ICS. The major objective of this focus group is to discuss and identify contributory factors and observations (or test results) for the problem which we considered. Each participant was provided with a set of questions as shown in Additional file 1: Appendix A. Most of these questions are open-ended that ask for factors that would contribute to the major cause of the considered problem (attack/technical failure) and tests that would provide additional information to distinguish between the major cause of the considered problem (attack/technical failure) after the problem is observed by the floodgate operator. For instance, we considered the problem *"the sensor sends incorrect water level measurements"* and asked the participants: *"Which contributory factors would increase the likelihood of the problem due to (accidental) sensor failure?"*. The moderator explained each question to the participants and facilitated the discussion among the participants to identify a set of contributory factors and observations (or test results) for the observable problem which we considered.

After the focus group workshop, we employed a questionnaire to gather data for constructing the qualitative BN model. We employed snowball sampling to recruit other participants for this study through initial participants. This sampling technique is useful as it helps to find experts in ICS safety and/or security quickly. The participants were provided with the same set of questions which we provided to focus group participants as shown in Additional file 1: Appendix A to elicit contributory factors and observations (or test results) for the considered problem. We received 10 responses in total for the questionnaire. However, we excluded one response as the participant did not have any experience working with ICS. Importantly, seven out of nine respondents have five or more years working experience with ICS which helps to ensure reliability of data. In addition, we had a good mix of participants from safety and/or security community which is important for our application. Specifically, two out of nine respondents associate themselves with both safety and security, two out of nine respondents associate themselves with safety and five out of nine respondents associate themselves with security.

We combined the data gathered from the focus group and questionnaire for coding. We utilised thematic coding by grouping contributory factors which are similar under a category. For instance, there were nine responses such as "easy access to sensor", "attacker has physical access to the sensor", "free access to sensor" which we categorised into "easy physical access to sensor". On the other hand, we grouped and removed contributory factors which are not contributory factors based on our definition. For instance, "Man-in-the-Middle attack using the wired connection" is not a specific contributory factor but rather a type of attack that an attacker might employ. Once we categorised the contributory factors, there were 14 categories (parent nodes) in total. However, this would result in the CPT size of the problem variable as 16,384, which makes it unmanageable. Therefore, we utilised parent node divorcing, which allows parent nodes to be grouped hierarchically to avoid excessive inbound links to the child node. By utilising parent node divorcing, we reduced the number of parent nodes to eight which in turn reduced the CPT size of the problem variable to 256. For instance, we grouped hierarchically three different parent nodes (location of sensor susceptible to severe weather, location of sensor susceptible to biological fouling, location of sensor susceptible to physical contact of marine vessel) into a single parent node (location of sensor susceptible to external factor) as shown in Fig. 4, because they are of the same theme and no original interactions are lost in the process. Figure 4 shows two different types of causal interactions between an individual contributory factor and the problem: (i) cause and (ii) inhibitor. The contributory factors including easy physical accessibility, no sensor data integrity verification, easy network accessibility, presence of software in sensor and no sensor firmware update have a positive influence on the problem (major cause for sensor sends incorrect water level measurements). On the other hand, the contributory factors including maintenance issue, no use of Electro-Magnetic Interference (EMI) shielding technique, location of sensor susceptible to external factor have an inhibiting influence on the problem. Once the qualitative

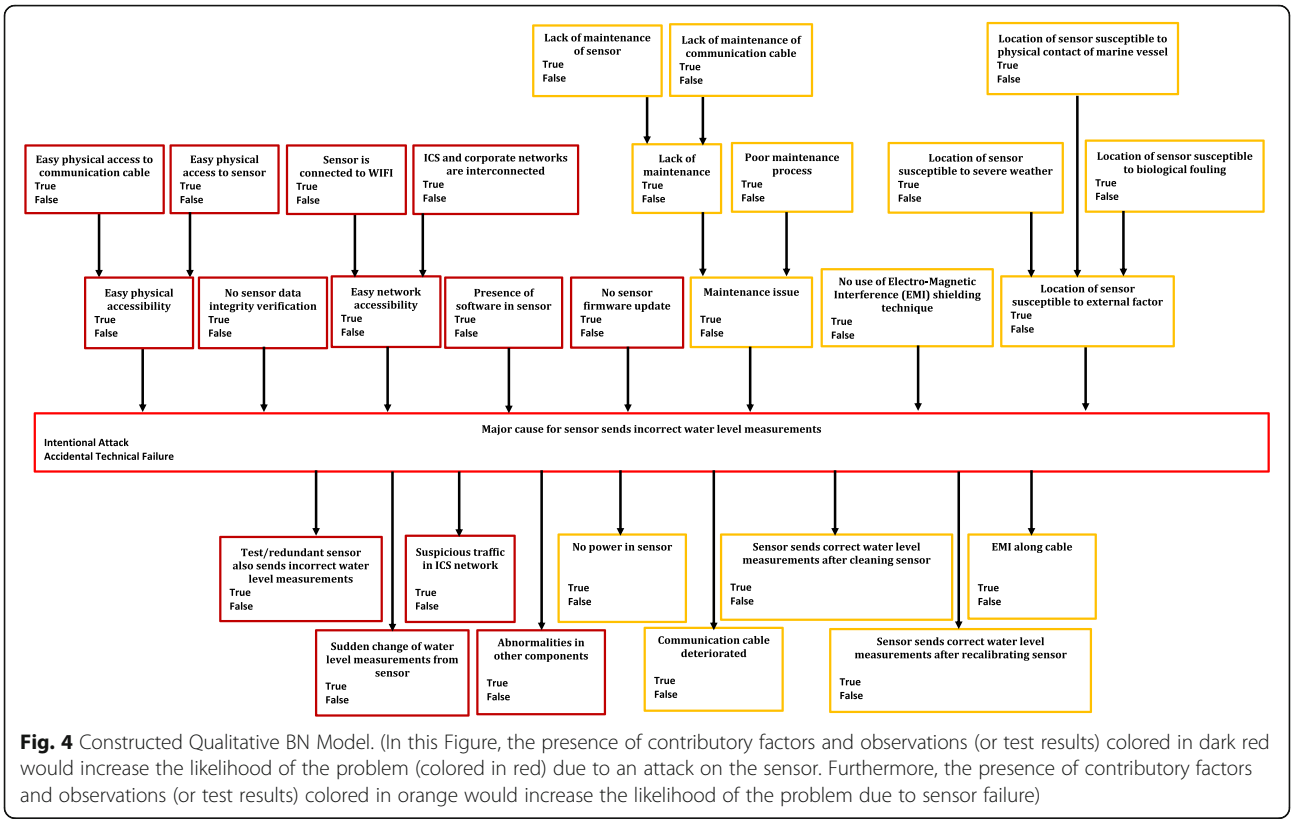

Fig. 4 Constructed Qualitative BN Model. (In this Figure, the presence of contributory factors and observations (or test results) colored in dark red would increase the likelihood of the problem (colored in red) due to an attack on the sensor. Furthermore, the presence of contributory factors and observations (or test results) colored in orange would increase the likelihood of the problem due to sensor failure)

BN model shown in Fig. 4 is constructed, we validated it through a focus group workshop with five experts who have experience working with safety and/or security of ICS in the water management sector in the Netherlands. We asked specifically whether anything is missing or not appropriate in the qualitative BN model. However, the experts did not find any need to add or update anything in the constructed qualitative BN model.

Construction of quantitative BN model for distinguishing attacks and technical failures in floodgates

A multimethodology approach is used for quantitative data collection like we did for the construction of the qualitative BN model. In order to gather data for populating the BN model with probabilities, we utilised a focus group workshop and a questionnaire. Firstly, we conducted a focus group workshop with five participants who have experience working with safety and/or security of ICS in the water management sector in the Netherlands. The major objective of this focus group is to elicit probabilities corresponding to each variable in our qualitative BN model that could help to determine the major cause (intentional attack or accidental technical failure) of the problem (sensor sends incorrect water level measurements) when observed.

Additional file 1: Appendix B shows a set of questions which we provided to each participant at the start of the focus group workshop. We asked each participant to answer the question using a probability scale with numerical and verbal anchors to elicit prior probabilities corresponding to the contributory factors and conditional probabilities corresponding to the problem and observations (or test results). For instance, we elicited the prior probability of the variable *"Easy Physical Access to Sensor"* and a conditional probability of the variable "Major cause for sensor sends incorrect water level measurements" as shown in Fig. 5. We utilised a systematic and effective way of eliciting marginal and conditional probabilities from the experts by asking appropriate types of questions, taking into account the type of causal interaction as shown in Table 1. It is evident that providing the fragment of text (i.e., the elicitation question) as shown in Figs. 5 and 6 instead of the mathematical notation to elicit conditional probabilities from domain experts worked very well on the development of a BN model supporting patient-specific therapy selection for oesophageal carcinoma (Van der Gaag et al. 2002). Furthermore, this is also employed in the development of a BN model using domain experts that help to assess the potential effects of establishing the ENSI navigation service to ship collisions and groundings (Hänninen et al. 2014). In our application, the participants were asked to answer the questions individually to avoid bias in their

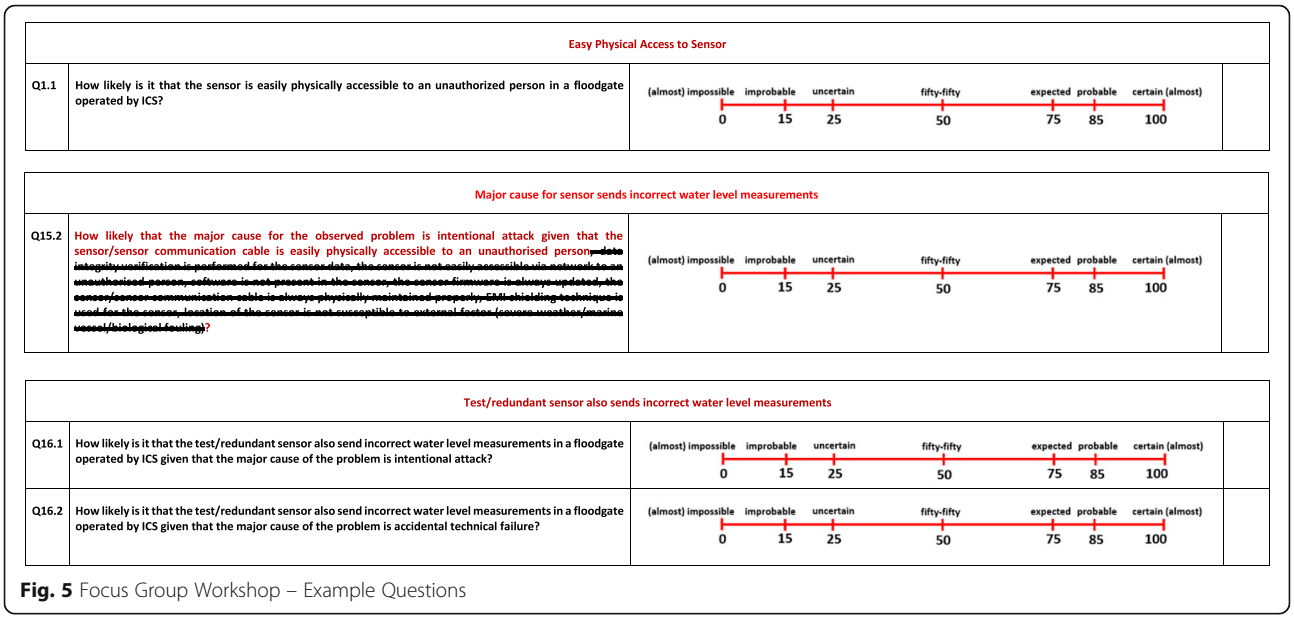

Fig. 5 Focus Group Workshop – Example Questions

responses. Furthermore, the moderator provided clarifications individually in case there are any questions from the participants. Once the participants answered the questions individually, the moderator facilitated a discussion on the reasoning behind the varied probabilities which they provided for some variables. However, the purpose of this discussion is not to make them reach a consensus as it could make the responses biased.

In addition to the focus group workshop, we utilised a questionnaire to gather data for populating the BN model with probabilities. We used snowball sampling to recruit other participants for this study through initial participants in the focus group workshop as the target group is limited and rare to find. This sampling technique makes it easier to find experts in safety and/or security of ICS in the water management sector in the Netherlands quickly. We provided a set of questions to the participants mainly to elicit probabilities corresponding to each variable in the constructed BN model as shown in Additional file 1: Appendix B. For instance, we asked for the prior probability of the variable *"Easy Physical Access to Sensor"* and a conditional probability of the variable "Major cause for sensor sends incorrect water level measurements" as shown in Fig. 6. The difference compared to the focus group workshop questions is that the probability scale with numerical and verbal anchors is not directly used as it is not practicable in the online questionnaire. However, we utilised the verbal and corresponding numerical anchors from the probability scale as answer choices for each question in the online questionnaire in addition to "others" option which could help participants to provide fine-grained probabilities as shown in Fig. 6. We received five responses in total. Overall, seven out of 10 participants have more than 5 years work experience with safety and/or security of ICS in the water management sector in the Netherlands.

Once we collected the responses from the participants in both the focus group workshop and questionnaire, we tabulated them together. Furthermore, we noticed that there were some missing data due to no or invalid response from some respondents. For instance, we considered responses like "others" without mentioning any specific likelihood value as an invalid response. Furthermore, it is also not possible to clarify with the respondent as responses are anonymous. Ignoring or discarding missing data is one of the most common approaches used to deal with the missing data (Baraldi and Enders 2010; Twala 2009). Listwise deletion and pairwise deletion are the two different methods which could help to ignore or discard the missing data (Baraldi and Enders 2010). Pairwise deletion is appropriate for our application as it ignores or discards only the missing data and considers the other data provided by these experts. This is easy to implement. Therefore, we utilised pairwise deletion to ignore or discard the missing data in our application. Listwise deletion is not appropriate for our application as it leads to loss of data by completely ignoring or discarding data from four out of 10 experts since they have no or invalid response to a question.

Once the missing data is ignored or discarded, the probabilities $P_i(X)$ elicited from the experts need to be combined. One of the most widely used method to combine the probabilities elicited from the experts is linear pooling (Farr et al. 2018; Ouchi 2004). Using the linear

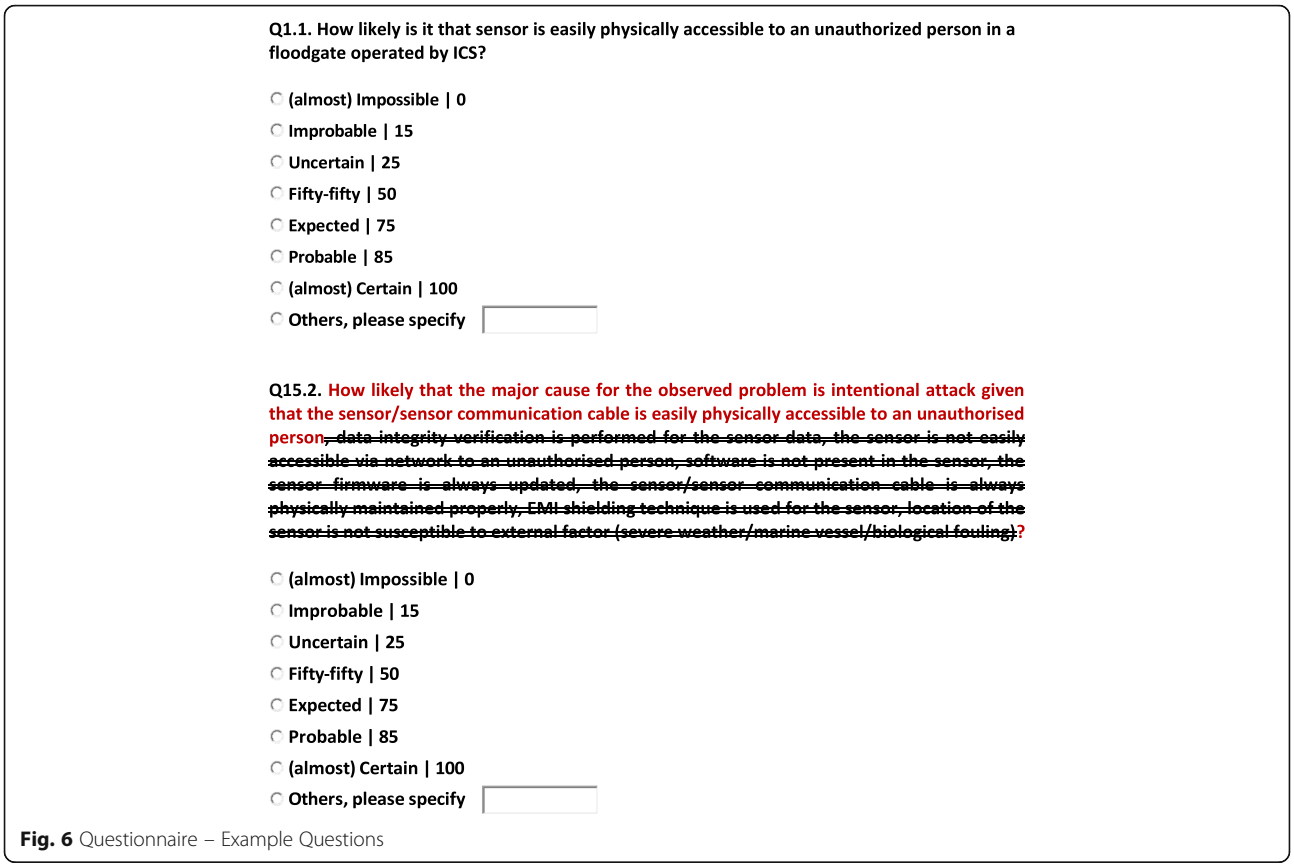

Fig. 6 Questionnaire – Example Questions

pooling method, the combined probabilities can be computed using (2):

$$P(X) = \sum_{i=1}^{n} w_i P_i(X) \tag{2}$$

Where w_i are positive weights given to each of the n experts with complete probabilities for the corresponding X and $\sum_{i=1}^{n} w_i = 1$.

There are two different types of linear pooling method: (i) prior linear pooling, and (ii) posterior linear pooling (Farr et al. 2018). Prior linear pooling combines elicited probabilities from experts corresponding to each variable in the BN model, which could then be used to compute posterior probabilities of target variables by providing evidences to some variables. On the other hand, in posterior linear pooling, elicited probabilities from n experts are used to construct n distinct BNs. Once we construct the n distinct BNs, we run these BNs by providing same evidences to the same set of variables in these BNs and compute different posterior probabilities in each of these BNs. Finally, the posterior probabilities generated in n distinct BNs are combined. However, this is not appropriate for our application as it is not practicable for performing diagnostics in a timely way. Furthermore, this is not suitable for our application as we ignored or discarded missing data which could make it not possible to construct BNs with no probabilities for some variables.

In our application, we utilised prior linear pooling as it is appropriate based on its advantages (Farr et al. 2018). Each of the 10 experts is given equal weighting as they all have experience working with safety and/or security of ICS in the water management sector in the Netherlands. Furthermore, we consider each respondent's experience to be equal in value. So, we combined the probabilities from n experts using (2).

The probabilities corresponding to contributory factors and observations (or test results) are now complete. However, we utilised DeMorgan model to reduce the number of CPT entries that needs to be elicited from experts to nine. Therefore, we computed the remaining CPT entries corresponding to the problem variable using (1). An excerpt of CPT entries corresponding to the problem variable is shown in Table 2. The complete BN model with both the qualitative and quantitative component is shown Fig. 7.

The DeMorgan model is not applicable for eliciting conditional probabilities of observations (or test results) as they only have one parent (i.e., the problem variable). Therefore, we elicited these probabilities directly from experts as they are straightforward. This is because the

CPT size of each observation (or test result) is 4 (2^{1+1}) as they only have one parent. For instance, we asked for a conditional probability of the variable "test/redundant sensor also sends incorrect water level measurements" taking into account the major cause ("*Intentional attack*"/"*Accidental technical failure*") of the observed problem ("*Sensor sends incorrect water level measurements*") is already known as shown in Fig. 5.

Demonstration of the constructed BN model

In this section, we demonstrate the suitability of the constructed BN model based on two different illustrative scenarios. It is not possible to utilise the real floodgate for demonstrating the suitability of the constructed BN model by putting it into practice due to availability and criticality issues. Therefore, we relied on two different illustrative scenarios for this purpose.

These two different illustrative scenarios help to show when and how the constructed BN model using the attack-failure distinguisher framework would be useful in practice. Firstly, we assume that the floodgate operator observed that a sensor sends incorrect water level measurements by noticing the mismatch between the measurements from physical water level scale and water level sensor. In order to choose the appropriate response strategy, the floodgate operator needs to determine the major cause of this problem (i.e., whether this problem is caused by an attack or technical failure), which is the aim of the constructed BN model.

Once the floodgate operator noticed the incorrect sensor measurements problem, they need to provide the evidence that is available for variables in the upper layer (contributory factors) and lower layer (test results). This could help the constructed BN model compute posterior probabilities of both the states in the problem variable (attack and technical failure) based on the provided evidences.

Table 2 CPT Excerpt – Problem Variable

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Y	
								Attack	Failure
True	True	True	True	True	True	True	True	0.02	0.98
True	True	True	True	True	True	True	False	0.09	0.91
True	True	True	True	True	True	False	True	0.06	0.94
True	True	True	True	True	True	False	False	0.24	0.76
True	True	True	True	True	False	True	True	0.09	0.91
True	True	True	True	True	False	True	False	0.38	0.62
True	True	True	True	True	False	False	True	0.24	0.76
True	True	True	True	True	False	False	False	0.97	0.03
True	True	True	True	False	True	True	True	0.02	0.98
True	True	True	True	False	True	True	False	0.09	0.91

In this table, C_1: Easy physical accessibility, C_2: No sensor data integrity verification, C_3: Easy network accessibility, C_4: Presence of software in sensor, C_5: No sensor firmware update, C_6: Maintenance issue, C_7: No use of EMI shielding technique, C_8: Location of sensor susceptible to external factor and Y: Major cause for sensor sends incorrect water level measurements

In the first illustrative scenario, the floodgate operator set evidence for variables based on the available information as shown in Table 3. Based on such evidence, the posterior probability is computed by the constructed BN model for other variables without any evidence. The BN model in Fig. 8 shows that the incorrect water level measurement problem is most likely due to technical failure based on the provided evidences. This information would help to select the appropriate response strategy (i.e., to repair or replace the water level sensor).

In the second illustrative scenario, the floodgate operator sets different evidence for variables in the constructed BN model based on the available information as shown in Table 3. Based on the provided evidences, the posterior probability is computed for other variables without any evidence in the constructed BN model. Figure 9 shows that the incorrect water level measurement problem is most likely due to attack based on the evidences provided by the floodgate operator. This information would help to choose the suitable response strategy (i.e., to block the corresponding attack vector).

The difference between the two scenarios can be explained as follows. In the first illustrative scenario, the sensor/sensor communication cable is not easily accessible to an unauthorised person, whereas there is a lack of maintenance of the sensor/sensor communication cable and the location of the sensor is susceptible to external factors such as biological fouling. In addition, the sensor communication cable is deteriorated, and the sensor sends correct water level measurements after cleaning the sensor. Typically, the above-mentioned factors increase the likelihood of the problem due to accidental technical failure, which is reflected in terms of the posterior probability of Y in Fig. 8. In contrast, in the second illustrative scenario, the sensor/sensor communication cable is properly maintained, and the location of the sensor is not susceptible to external factors such as biological fouling, whereas the sensor/sensor communication cable is easily physically accessible to an unauthorised person. In addition, the test/redundant sensor also sends incorrect water level measurements. Typically, the above-mentioned factors increase the likelihood of the problem due to intentional attack, which is reflected in terms of the posterior probability of Y in Fig. 9.

Discussion

This section mainly highlights and discusses implications, limitations of this study and potential future directions.

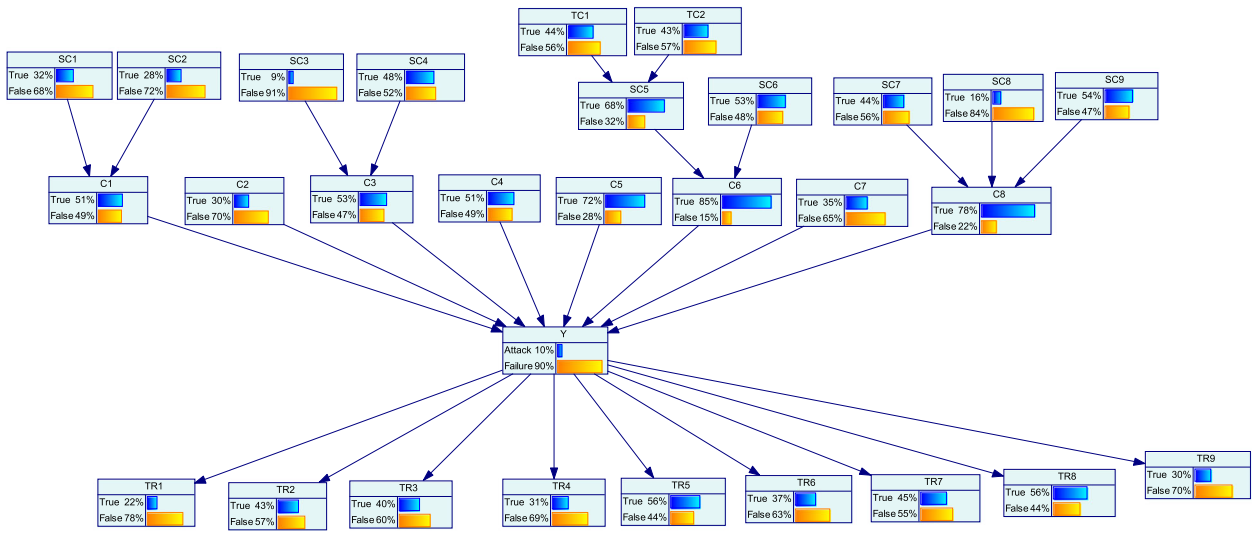

Fig. 7 Constructed BN Model – No Evidence Provided. (In this figure, *SC*1: Easy physical access to sensor, *SC*2: Easy physical access to communication cable, *C*1: Easy physical accessibility, *C*2: No sensor data integrity verification, *SC*3: Sensor is connected to WIFI, *SC*4: ICS and corporate networks are interconnected, *C*3: Easy network accessibility, *C*4: Presence of software in sensor, *C*5: No sensor firmware update, *TC*1: Lack of maintenance of sensor, *TC*2: Lack of maintenance of communication cable, *SC*5: Lack of maintenance, *SC*6: Poor maintenance process, *C*6: Maintenance issue, *C*7: No use of EMI shielding technique, *SC*7: Location of sensor susceptible to severe weather, *SC*8: Location of sensor susceptible to physical contact of marine vessel, *SC*9: Location of sensor susceptible to biological fouling, *C*8: Location of sensor susceptible to external factor, *Y*: Major cause for sensor sends incorrect water level measurements, *TR*1: Test/redundant sensor also sends incorrect water level measurements, *TR*2: Sudden change of water level measurements from sensor, *TR*3: Suspicious traffic in ICS network, *TR*4: Abnormalities in other components, TR5: No power in sensor, *TR*6: Communication cable deteriorated, *TR*7: Sensor sends correct water level measurements after cleaning sensor, *TR*8: Sensor sends correct water level measurements after recalibrating sensor, *TR*9: EMI along cable)

The results of existing integrated safety and security risk assessment methods would help to choose suitable risk treatments during the design phase before an attack or technical failure occurs. On the other hand, our method involving the attack-failure distinguisher framework would help to choose appropriate response strategies during the operational phase when an attack or technical failure occurs. Furthermore, our method would help operators to think more proactively about reactive safety and security.

As a part of the probability elicitation process, in addition to the case outline, we also provided information related to the type of floodgate (criticality rating: very high) and context (threat level: substantial). This guideline helps to avoid very diverse responses over participants as they have substantive information based on the system knowledge. The raw data on elicited probabilities is not shared due to criticality and sensitivity issues. However, we provided boxplots based on probabilities elicited from experts for five variables as shown in Fig. 10. This shows the diversity of raw data on elicited probabilities. In particular, the interquartile range for most of these variables are low which confirms that raw data on elicited probabilities is less dispersed. This also indicates that experts have more or less common understanding of the system with the limited system information provided as we relied on experts who have experience working with safety and/or security of ICS in the water management sector in the Netherlands. However, the elicited probabilities can be further refined by providing additional system information details to make it more realistic in the future as the CPT values highly likely depend on the specifics of a particular system.

We constructed the qualitative BN model using the data gathered from a focus group workshop and questionnaire that had 14 participants in total. This focus group workshop included five participants who are experts on safety and/or security of ICS in the water management sector in the Netherlands. Furthermore, this was complemented with a questionnaire which had nine respondents who have at least a year of experience working with safety and/or security of ICS in general from different countries. Finally, the constructed qualitative BN model was validated through a focus group workshop, which had the participation of five experts on safety and/or security of ICS in the water management sector in the Netherlands. In terms of generalisability, the constructed qualitative BN model can be used as a starting point for constructing a BN model for the same problem (incorrect sensor measurements) in a similar type of infrastructure in a different country or in another sector. This can be further updated and validated by involving experts in the corresponding country/sector.

Table 3 Evidences Corresponding to both the Illustrative Scenarios

Name of the Variable	Evidences (First Illustrative Scenario)	Evidences (Second Illustrative Scenario)
Easy physical access to sensor ($SC1$)	False	True
Easy physical access to communication cable ($SC2$)	False	True
No sensor data integrity verification ($C2$)	False	True
Sensor is connected to WIFI (SC3)	False	False
ICS and corporate networks are interconnected ($SC4$)	False	False
Presence of software in sensor ($C4$)	False	True
No sensor firmware update ($C5$)	False	True
Lack of maintenance of sensor ($TC1$)	True	False
Lack of maintenance of communication cable ($TC2$)	True	False
Poor maintenance process ($SC6$)	True	False
No use of EMI shielding technique ($C7$)	False	True
Location of sensor susceptible to severe weather ($SC7$)	True	False
Location of sensor susceptible to physical contact of marine vessel ($SC8$)	True	False
Location of sensor susceptible to biological fouling ($SC9$)	True	False
Test/Redundant sensor also sends incorrect water level measurements ($TR1$)	False	True
Sudden change of water level measurements from sensor ($TR2$)	False	False
Suspicious traffic in ICS network ($TR3$)	True	True
Abnormalities in other components ($TR4$)	True	True
Communication cable deteriorated ($TR6$)	True	True
Sensor sends correct water level measurements after cleaning sensor ($TR7$)	True	False

On the other hand, we constructed the quantitative BN model based on elicited probabilities through a focus group workshop and questionnaire. The focus group workshop included five experts on safety and/or security of ICS in the water management sector in the Netherlands. Furthermore, this was complemented with a questionnaire which had five respondents who have at least a year of experience working with safety and/or security of ICS in water management sector in the Netherlands. During these elicitation processes, in addition to the case outline, we also provided information corresponding to the type of floodgate and context which includes the criticality rating and threat level. With regard to generalisability of the quantitative BN model, this is limited as it is only applicable to a specific type of floodgates in a specific context in the Netherlands. Furthermore, this is also not directly generalisable to the same problem (incorrect sensor measurements) in a similar type of infrastructure in a different country or in another sector. Therefore, the probabilities need to be elicited from experts in the corresponding country/domain directly to make the quantitative BN model reliable.

There are two key limitations which impacted our sample size for focus group workshops and questionnaire: (i) Limited experts on safety and/or security of ICS in the water management sector: We relied on experts who associate themselves with safety and/or security of ICS to elicit contributory factors and test results (or observations). We also relied on experts who associate themselves with safety and/or security of ICS in the water management sector in the Netherlands to elicit

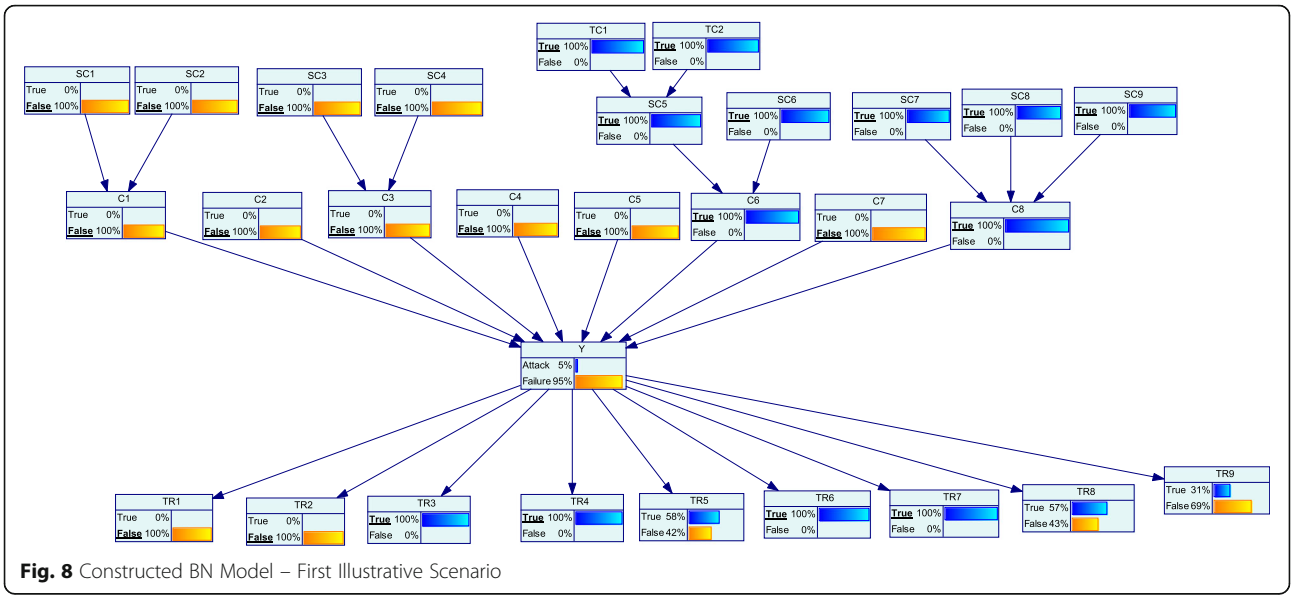

Fig. 8 Constructed BN Model – First Illustrative Scenario

probabilities. This enhances the reliability of elicited contributory factors, test results (or observations) and probabilities as they have prior knowledge about the system. However, this leads to the limitation of fewer respondents. In the Netherlands, there is a limited group of safety and/or security experts in the water management sector. Therefore, we utilised snowball sampling as it helps to reach more number experts in that limited target group, (ii) Limited time availability of experts: Initially, we employed focus groups as a technique to elicit contributory factors, test results (or observations) and probabilities. However, there were practical difficulties to gather a group of people at the same time due to the limited time availability of experts. This resulted in focus groups with a somewhat lower number of experts (five). Therefore, we complemented focus groups with questionnaires to reach a bit more number of experts in that limited target group. Due to limited target group and time availability of experts, it was not possible to reach much more experts to elicit contributory factors, test results (or observations) and probabilities.

However, due to such limitations, it seems to be prevalent in practice to have a group size less than 10 (Hänninen et al. 2014; Van der Gaag et al. 2002). For instance, eight experts with maritime working experience helped in the construction of BN model to assess the potential effects of establishing the ENSI navigation service to ship collisions and groundings. Furthermore, in an another

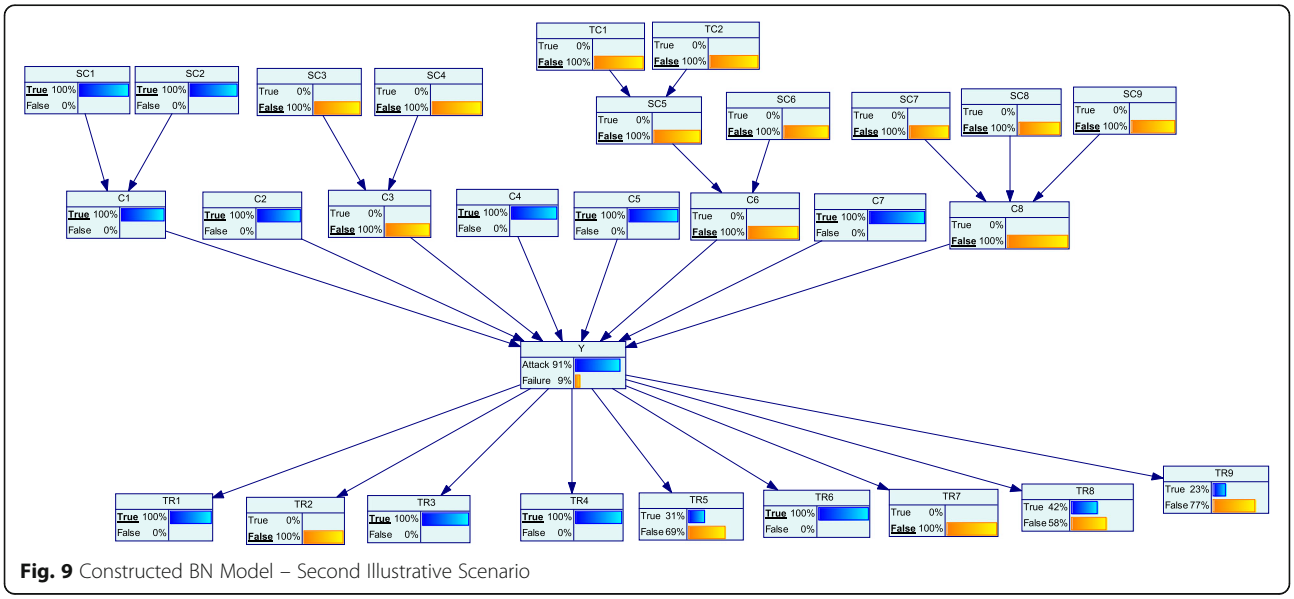

Fig. 9 Constructed BN Model – Second Illustrative Scenario

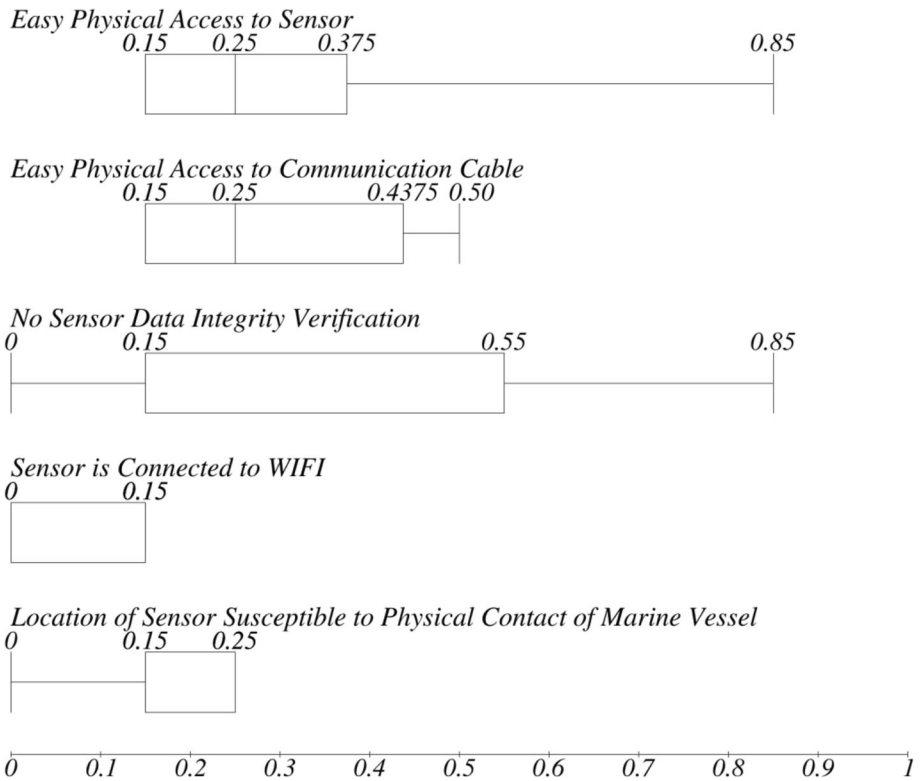

Fig. 10 Boxplots for Different Variables in the Constructed BN Model

instance, two experts in gastrointestinal oncology helped in the development of decision support for patient-specific therapy selection for oesophageal carcinoma (Van der Gaag et al. 2002).

In this paper, we provided a case study of attack-failure distinguisher framework by developing a BN model for the problem of incorrect sensor measurements in floodgates. Furthermore, we provided two different illustrative scenarios using the developed BN model to demonstrate the suitability of such models. In the future, this would help practitioners to develop BN models for different problems in different sectors. When we rely on expert knowledge as the data source, there need to be appropriate methods to effectively elicit knowledge from experts. The attack-failure distinguisher framework includes extended fishbone diagrams to support brainstorming with experts in constructing the DAGs of BN models for our application. Furthermore, the attack-failure distinguisher framework includes DeMorgan model and probability scale with numerical and verbal anchors to effectively elicit probabilities from experts to completely define CPTs of BN models for our application. Some of these methods have already been applied separately in practical applications for different problems in different sectors (Hänninen et al. 2014; Jacobs 2018; Van der Gaag et al. 2002). For instance, Jacobs used extended fishbone diagrams for an example ProRail case related to carriage registration (Jacobs 2018). They populated the contributory factors on the left side of the extended fishbone diagram for the problem ("Incorrect registration"). Furthermore, they populated with observations on the right side of the extended fishbone diagram. This shows that individual components of the attack-failure distinguisher framework are effective and applicable for different problems in different sectors.

Historical data on attacks and technical failures in the water management sector in the Netherlands is unavailable for research due to sensitivity issues. Therefore, it would not be possible to develop models that could help to distinguish between attacks and technical failures for the problem of incorrect sensor measurements using a data-driven approach. However, in the future, the unavailability of historical data on attacks and technical failures would not deter modelling cyber security for ICS anymore as we utilised a knowledge-based approach to develop a model for distinguishing attacks and technical failures.

The other alternate data sources such as red team vs. blue team exercises were not possible due to practicalities, especially there is a lack of testbeds which could facilitate such exercises in the Netherlands. Such data

sources could further improve the reliability of data used to construct DAG and populate CPTs. Notably, the Critical Infrastructure Security Showdown (CISS) is conducted by Singapore University of Technology and Design on their Secure Water Treatment (SWaT) testbed (Antonioli et al. 2017). Such type of events could provide information about contributory factors and observations (or test results) corresponding to attacks. For instance, we could interview members of the red team regarding which factors in the infrastructure contributed to the success of their attack. Furthermore, we could interview members of the blue team regarding tests (or observations) which helped them to diagnose an attack. The use of existing testbeds like SWaT testbed is not appropriate for this study as it did not reflect the system which we considered i.e., a specific type of floodgates in the Netherlands. Therefore, there is a need for a testbed in the Netherlands which reflect the system which we considered for using red team vs. blue team exercises as an alternate data source and/or a system for evaluation in the future.

The real water management infrastructure like a floodgate is not available for the evaluation of the developed BN model due to availability and criticality issues. Therefore, we could not perform naturalistic evaluation, which involves evaluating the developed artefact with real users and real systems in the real setting. Therefore, we relied on the artificial evaluation, which involves evaluating the developed BN model in a contrived and non-realistic way. However, we made it more realistic with real-users, and realistic problems to correspond the results to real use. Furthermore, the developed BN model is validated using expert evaluation and illustrative scenarios. However, the quantitative BN model needs to be further evaluated using a testbed in the future. Currently, this was not possible due to the lack of testbed in the Netherlands which reflect the system which we considered. However, this evaluation would also help to answer the key question on how much confidence should an operator have based on such BN-based analysis.

Related work

This section highlights application of BNs in different domains. Furthermore, we summarise important patterns corresponding to the application of BNs in cyber security, which we used as a basis to develop BN models for our application. In addition, we point out studies that relate to the problem of distinguishing attacks and technical failures.

BNs are used for developing medical decision support systems (Curiac et al. 2009; Kahn et al. 2001; Kahn Jr et al. 1997; Luciani et al. 2003; Milho and Fred 2001; Onisko et al. 1999). Furthermore, BNs are also used in fault diagnosis (Cai et al. 2014; Huang et al. 2008; Zhao et al. 2013), cyber security (Alile 2018; Apukhtin 2011; Axelrad et al. 2013; Elmrabit et al. 2020; Greitzer et al. 2010; Greitzer et al. 2012; Herland et al. 2016; Holm et al. 2015; Ibrahimović and Bajgorić 2016; Kornecki et al. 2013; Kwan et al. 2009; Kwan et al. 2008; Mo et al. 2009; Pappaterra 2021; Pecchia et al. 2011; Shin et al. 2015; Wang and Guo 2010; Zhou et al. 2018).

In our previous work, we conducted a systematic literature review of BN models in cyber security (Chockalingam et al. 2017). In that study, we identified 17 standard BN models in cyber security based on the review methodology we adopted. The identified BN models were analysed using eight different criteria: (i) citation details, (ii) data sources used to construct Directed Acyclic Graphs (DAGs) and populate Conditional Probability Tables (CPTs), (iii) the number of nodes used in the model, (iv) type of threat actor, (v) application and application sector, (vi) scope of variables, (vii) the approach(es) used to validate models and (viii) model purpose and type of purpose.

Some of the important patterns in the use of standard BN models in cyber security which we identified includes: (i) data sources used to construct DAGs and populate CPTs in the identified BN models were expert knowledge and empirical data predominantly from cyber security reports, (ii) the identified BN models were predominantly used to tackle problems associated with the Information Technology (IT) environment compared to the ICS environment and (iii) the identified BN models completely or partially benefited risk management, forensic investigation, governance, threat hunting and vulnerability management in cyber security.

The identified BN models were considered as a starting point to develop the attack-failure distinguisher framework for constructing BN models that would help to distinguish between attacks and technical failures (Chockalingam et al. 2019). Furthermore, some of the identified patterns in the use of BN models in cyber security were used as a basis to construct BN models for our application. For instance, expert knowledge is a successful and well-established alternate data source to tackle problems associated with ICS environment as there is a no availability of data from real-world systems which we considered. Finally, some of the identified patterns in the use of BN models in cyber security were used as a motivation for this study to fill an identified research gap in addition to considering inputs from such BN models. For instance, we developed a BN model to tackle a problem associated with the ICS environment taking into account BN models used to tackle problems in IT environment.

(Ahmed et al. 2020) highlighted that distinguishing attacks and technical failures is necessary based on

interviews with researchers at state-of-the-art testbeds like SWaT, ICS security experts and engineers at industrial production plants for steel and water. Furthermore, they described three important challenges of distinguishing attacks and technical failures. One of the challenges is that related works mainly focus on the consequences of an attack or technical failure instead of looking into the properties of an attack or technical failure. Furthermore, they suggested different potential directions that could help to tackle the problem of distinguishing attacks and technical failures, one of which is to use data from both the network layer and the process layer.

There are a lot of works that focus on either detecting an attack or a technical failure separately. For instance, (Park et al. 2015) proposed an approach to detect sensor attacks in the presence of transient faults like a GPS reporting incorrect measurements inside a tunnel. Furthermore, (Samara et al. 2008) proposed a method for detection of sensor abrupt faults. However, these lack the capability to distinguish between attacks and technical failures.

Finally, (Anwar et al. 2015) proposed a data-driven approach to distinguish cyber-attacks from physical faults in a smart grid. Furthermore, they compared their approach with the conventional supervised classification approaches. However, their approach is not applicable when there is a lack of data which is typically the case in cyber security of different domains like water management.

Conclusions and future work

Harmful consequences of a problem could be minimised by choosing the appropriate response strategy in a timely manner. However, this is not possible without determining the major cause of a problem. In our previous work, we developed the attack-failure distinguisher framework which could help to construct BN models that determine whether the problem is caused by an attack or technical failure. This framework also includes the knowledge elicitation methods such as the DeMorgan model, and probability scales with numerical and verbal anchors to effectively elicit expert knowledge to construct such BN models. This work mainly focused on providing a full case study of the framework on how to construct the BN model for a problem and demonstrate when and how this could be used in practice.

In this work, we developed a BN model for the problem of incorrect sensor measurements in floodgates in the Netherlands using the attack-failure distinguisher framework. This corresponds to the second main contribution of this paper. Due to the lack of data, we relied on expert knowledge to construct the qualitative and quantitative part of the BN model for our problem. We elicited contributory factors and test results (or observations) through a focus group workshop and a questionnaire among respondents who have experience working with ICS. The data from both the focus group workshop and questionnaire were used to construct the qualitative BN model, which was also validated with five experts.

Once the qualitative BN model was constructed, we used the DeMorgan model to reduce the number of CPT entries that needs to be elicited for the problem variable to nine instead of 256. Firstly, we elicited probabilities corresponding to contributory factors, problem and test results (or observations) from experts who have experience working with safety and/or security of water management infrastructures operated by ICS in the Netherlands through a focus group workshop and questionnaire. During this elicitation, we employed probability scales with numerical and verbal anchors to facilitate individual probability entry by providing it as a visual aid. We computed the rest of the probabilities for the problem variable using the DeMorgan model. The process of using attack-failure distinguisher framework to construct the BN model for our application relates to the first main contribution of this paper. Finally, we demonstrated the suitability of the constructed BN model using two different illustrative scenarios. This associates with the third main contribution of this paper. The first illustrative scenario shows that the most likely cause for the considered problem is technical failure, whereas the second illustrative scenario shows that the most likely cause for the considered problem is attack based on the evidences provided.

It was not possible to use real systems for evaluating the attack-failure distinguisher framework due to availability and criticality issues. However, we utilised real-users and realistic problems to evaluate the attack-failure distinguisher framework by developing a prototype and using the developed prototype for two different illustrative scenarios to relate the results to real use. Therefore, the developed BN model is usable in real settings in the future. However, this BN model can be further updated with appropriate contributory factors, test results and probabilities based on the performance measures in the confusion matrix, which includes four different combinations of diagnosed and actual classes. This is only possible when a dataset corresponding to the problem in the real setting is available for research.

In the future, it would be beneficial to put the constructed BN model into practice in a real floodgate in case it is available to showcase the value of the constructed BN model. Furthermore, we developed a root-cause analysis framework with the appropriate type of variables and relationships between them in our previous work, which would help to construct BN models to

determine the attack-vector (in case of an attack) and failure mode (in case of a technical failure) (Chockalingam and Katta 2019). However, the root-cause analysis framework needs to be applied and evaluated for a problem like incorrect sensor measurements in the future as it could complement the attack-failure distinguisher framework to determine the attack-vector (in case of an attack) and failure mode (in case of a technical failure). This could also help to choose the most effective response strategy between alternatives like repairing or replacing the sensor.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42400-021-00086-6.

Additional file 1 : **Appendix A**. Knowledge Elicitation Method to Develop Qualitative BN Model. **Appendix B**. Knowledge Elicitation Method to Develop Quantitative BN Model.

Acknowledgements

The authors would like to thank the focus group participants and questionnaire respondents. We also thank the anonymous reviewers for their time and efforts in reviewing our manuscript and providing constructive comments.

Authors' contributions

Sabarathinam Chockalingam: Conceptualization, Methodology, Investigation, Validation, Writing – Original Draft. Wolter Pieters: Writing – Review & Editing, Supervision. André Teixeira: Writing – Review & Editing, Supervision. Pieter van Gelder: Writing – Review & Editing, Supervision. All authors read and approved the final manuscript.

Funding

This research received funding from the Netherlands Organization for Scientific Research (NWO) in the framwork of the Cyber Security research program under the project "Secure Our Safety: Building Cyber Security for Flood Management (SOS4Flood)".

Availability of data and materials

The probabilities elicited from experts during this research will not be shared, due to the data criticality and sensitivity issues. The elicited probabilities are critical and sensitive because they are elicited from experts who have experience working with safety and/or security of ICS in the water management sector in the Netherlands, which reflects reality for a specific type of floodgates under a particular threat level. However, corresponding excerpts are provided in the paper to make application of BNs for distinguishing attacks and technical failures comprehensible, which also ensures that the value of this research is not negatively impacted by not sharing the elicited probabilities.

Declaration

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands. [2]Department of Risk, Safety and Security, Institute for Energy Technology, Halden, Norway. [3]Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands. [4]Department of Electrical Engineering, Uppsala University, Uppsala, Sweden.

Received: 12 August 2020 Accepted: 4 April 2021
Published online: 01 September 2021

References

Ahmed, C. M., Prakash, J., & Zhou, J. (2020). Revisiting anomaly detection in ICS: aimed at segregation of attacks and faults. *arXiv preprint arXiv:2005.00325*

Alile OS (2018) Predicting multi-stage attack with normal IP addresses on a computer network using Bayesian belief network. University of Benin, Benin

Antonioli D, Ghaeini HR, Adepu S, Ochoa M, Tippenhauer NO (2017) Gamifying ICS security training and research: design, implementation, and results of S3. In: Proceedings of the Workshop on Cyber-Physical Systems Security and Privacy

Anwar A, Mahmood AN, Shah Z (2015) A data-driven approach to distinguish cyber-attacks from physical faults in a smart grid. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management

Apukhtin V (2011) Bayesian network modeling for analysis of data breach in a bank. University of Stavanger, Norway

Axelrad ET, Sticha PJ, Brdiczka O, Shen J (2013) A Bayesian network model for predicting insider threats. In: 2013 IEEE security and privacy workshops

Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48(1):5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Brewer J, Hunter A (1989) Multimethod research: a synthesis of styles. Sage library of social research (vol. 175). Sage Publications, Inc

Cai B, Liu Y, Fan Q, Zhang Y, Liu Z, Yu S, Ji R (2014) Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network. Appl Energy 114:1–9. https://doi.org/10.1016/j.apenergy.2013.09.043

Castellon N, Frinking E (2015) Securing critical infrastructures in the Netherlands: towards a National Testbed. The Hague Centre for Strategic Studies. Retrieved from https://www.thehaguesecuritydelta.com/media/com_hsd/report/53/document/Securing-Critical-Infrastructures-in-the-Netherlands.pdf

Chockalingam S, Katta V (2019) Developing a bayesian network framework for root cause analysis of observable problems in cyber-physical systems. In: 2019 IEEE Conference on Information and Communication Technology (CICT)

Chockalingam S, Pieters W, Teixeira A, Khakzad N, van Gelder P (2019) Combining Bayesian networks and fishbone diagrams to distinguish between intentional attacks and accidental technical failures. Graphical Models Secur (GramSec). https://doi.org/10.1007/978-3-030-15465-3_3

Chockalingam S, Pieters W, Teixeira A, van Gelder P (2017) Bayesian network models in cyber security: a systematic review. In: Nordic Conference on Secure IT Systems (NordSec). https://doi.org/10.1007/978-3-319-70290-2_7

Chockalingam S (2020) Distinguishing attacks and failures in industrial control systems: knowledge-based design of Bayesian networks for Water Management Infrastructures – Chapter 5 (Doctoral Thesis, Delft University of Technology, Delft, The Netherlands). Retrieved from https://doi.org/10.4233/uuid:17da1df4-3295-45d3-9119-9f92a547e7c6

Curiac D-I, Vasile G, Banias O, Volosencu C, Albu A (2009) Bayesian network model for diagnosis of psychiatric diseases. In: Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces

Darwiche A (2008) Bayesian networks. Foundations Artif Intell 3:467–509. https://doi.org/10.1016/S1574-6526(07)03011-8

Effendi A, Davis R (2015) ICS and IT: managing cyber security across the enterprise. In: SPE Middle East Intelligent Oil and Gas Conference and Exhibition

Elmrabit N, Yang S-H, Yang L, Zhou H (2020) Insider threat risk prediction based on Bayesian network. Comput Secur 96:101908. https://doi.org/10.1016/j.cose.2020.101908

Fallet-Fidry G, Weber P, Simon C, Iung B, Duval C (2012) Evidential network-based extension of leaky Noisy-OR structure for supporting risks analyses. In: Fault detection, supervision and safety of technical processes

Farr C, Ruggeri F, Mengersen K (2018) Prior and posterior linear pooling for combining expert opinions: uses and impact on Bayesian networks — the case of the wayfinding model. Entropy 20(3):209. https://doi.org/10.3390/e20030209

Greitzer FL, Kangas LJ, Noonan CF, Dalton AC (2010) Identifying at-risk employees: a behavioral model for predicting potential insider threats. Pacific Northwest National Lab (PNNL), Richland. Retrieved from https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-19665.pdf

Greitzer FL, Kangas LJ, Noonan CF, Dalton AC, Hohimer RE (2012) Identifying at-risk employees: modeling psychosocial precursors of potential insider threats. In: 2012 45th Hawaii International Conference on System Sciences

Hänninen M, Mazaheri A, Kujala P, Montewka J, Laaksonen P, Salmiovirta M, Klang M (2014) Expert elicitation of a navigation service implementation effects on ship groundings and collisions in the Gulf of Finland. Proc Inst Mech Eng, Part O: J Risk Reliability 228(1):19–28

Herland K, Hämmäinen H, Kekolahti P (2016) Information security risk assessment of smartphones using Bayesian networks. J Cyber Secur Mobility 4(3):65–86. https://doi.org/10.13052/jcsm2245-1439.424

Holm H, Korman M, Ekstedt M (2015) A Bayesian network model for likelihood estimations of acquirement of critical software vulnerabilities and exploits. Inf Softw Technol 58:304–318. https://doi.org/10.1016/j.infsof.2014.07.001

Holm H, Sommestad T, Ekstedt M, NordströM L (2013) CySeMoL: a tool for cyber security analysis of enterprises. In: 22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013)

Huang Y, McMurran R, Dhadyalla G, Jones RP (2008) Probability based vehicle fault diagnosis: Bayesian network method. J Intell Manuf 19(3):301–311. https://doi.org/10.1007/s10845-008-0083-7

Husák M, Komárková J, Bou-Harb E, Čeleda P (2018) Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun Surveys Tutorials 21(1):640–660

Ibrahimović S, Bajgorić N (2016) Modeling information system availability by using Bayesian belief network approach. Interdiscip Description Complex Syst: INDECS 14(2):125–138. https://doi.org/10.7906/indecs.14.2.2

Jacobs F (2018) Safety through machine learning applications: a safety case analysis (Master thesis, Delft University of Technology, Delft). Retrieved from http://resolver.tudelft.nl/uuid:ce5c73ef-8ad0-426f-926e-7d7ef3e197c3

Kahn CE, Laur JJ, Carrera G (2001) A Bayesian network for diagnosis of primary bone tumors. J Digit Imaging 14(1):56–57. https://doi.org/10.1007/BF03190296

Kahn CE Jr, Roberts LM, Shaffer KA, Haddawy P (1997) Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med 27(1):19–29. https://doi.org/10.1016/S0010-4825(96)00039-X

Knowles W, Prince D, Hutchison D, Disso JFP, Jones K (2015) A survey of cyber security management in industrial control systems. Int J Crit Infrastruct Prot 9:52–80. https://doi.org/10.1016/j.ijcip.2015.02.002

Kornecki AJ, Subramanian N, Zalewski J (2013) Studying interrelationships of safety and security for software assurance in cyber-physical systems: approach based on Bayesian belief networks. In: 2013 Federated Conference on Computer Science and Information Systems

Kraaijeveld P (2005) Genierate: an interactive generator of diagnostic Bayesian network models. In: 16th International Workshop on Principle Diagnosis

Kwan M, Chow K-P, Lai P, Law F, Tse H (2009) Analysis of the digital evidence presented in the yahoo! Case. In: IFIP International Conference on Digital Forensics

Kwan M, Chow K-P, Law F, Lai P (2008) Reasoning about evidence using Bayesian networks. In: IFIP International Conference on Digital Forensics

Luciani D, Marchesi M, Bertolini G (2003) The role of Bayesian networks in the diagnosis of pulmonary embolism. J Thromb Haemost 1(4):698–707. https://doi.org/10.1046/j.1538-7836.2003.00139.x

Maaskant PP, Druzdzel MJ (2008) An Independence of Causal Interactions Model for Opposing Influences. In: 4th European workshop on probabilistic graphical models

Martin TG, Burgman MA, Fidler F, Kuhnert PM, Low-Choy S, McBride M, Mengersen K (2012) Eliciting expert knowledge in conservation science. Conserv Biol 26(1):29–38. https://doi.org/10.1111/j.1523-1739.2011.01806.x

Milho I, Fred A (2001) A user-friendly development tool for medical diagnosis based on Bayesian networks. In: Sharp B, Filipe J, Cordeiro J (eds) Enterprise Information Systems II. Springer, Dordrecht, pp 113–118. https://doi.org/10.1007/978-94-017-1427-3_16

Mo SYK, Beling PA, Crowther KG (2009) Quantitative assessment of cyber security risk using Bayesian network-based Model. In: 2009 Systems and Information Engineering Design Symposium

Nakatsu RT (2009) Diagrammatic Reasoning in AI. Wiley, Hoboken. https://doi.org/10.1002/9780470400777

Nogueira HIS, Walraven M (2018) Overview storm surge barriers. Rijskwaterstaat, Deltares. Retrieved from http://www.masterpiece.dk/UploadetFiles/10852/25/Deltares_2018_Overview_storm_surge_barriers_komprimeret.pdf

Onisko A, Druzdzel MJ, Wasyluk H (1999) A Bayesian network model for diagnosis of liver disorders. In: Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering

Ouchi F (2004) A literature review on the use of expert opinion in probabilistic risk analysis. World Bank Policy Research Working Paper 3201

Pappaterra MJ, Flammini F (2021) Bayesian networks for online cybersecurity threat detection. In: Machine intelligence and big data analytics for cybersecurity applications (pp. 129-159). Springer, Cham. https://doi.org/10.1007/978-3-030-57024-8_6

Park J, Ivanov R, Weimer J, Pajic M, Lee I (2015) Sensor attack detection in the presence of transient faults. In: Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems

Pecchia A, Sharma A, Kalbarczyk Z, Cotroneo D, Iyer RK (2011) Identifying compromised users in shared computing infrastructures: a data-driven Bayesian network approach. In: 2011 IEEE 30th International Symposium on Reliable Distributed Systems

RISI. (2014). German steel mill cyber attack. Retrieved from http://www.risidata.com/database/detail/german-steelmill-cyber-attack.

Samara PA, Fouskitakis GN, Sakellariou JS, Fassois SD (2008) A statistical method for the detection of sensor abrupt faults in aircraft control systems. IEEE Trans Control Syst Technol 16(4):789–798. https://doi.org/10.1109/TCST.2007.903109

Shin J, Son H, Heo G (2015) Development of a cyber security risk model using Bayesian networks. Reliability Eng Syst Saf 134:208–217. https://doi.org/10.1016/j.ress.2014.10.006

Skopik F, Smith PD (eds) (2015) Smart grid security: innovative solutions for a modernized grid. Syngress, Boston. https://doi.org/10.1016/C2014-0-01356-1

Twala B (2009) An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell 23(5):373–405. https://doi.org/10.1080/08839510902872223

Van der Gaag LC, Renooij S, Witteman CL, Aleman BM, Taal BG (2002) Probabilities for a probabilistic network: a case study in Oesophageal cancer. Artif Intell Med 25(2):123–148. https://doi.org/10.1016/S0933-3657(02)00012-X

Wang JA, Guo M (2010) Vulnerability categorization using Bayesian networks. In: Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research

Zagorecki A (2010) Local probability distributions in Bayesian networks: knowledge elicitation and inference (Doctoral Dissertation, University of Pittsburgh, Pittsburgh). Retrieved from http://d-scholarship.pitt.edu/6542/1/Zagorecki_April_21_2010.pdf

Zhang G, Thai VV (2016) Expert elicitation and Bayesian network modeling for shipping accidents: a literature review. Saf Sci 87:53–62. https://doi.org/10.1016/j.ssci.2016.03.019

Zhao Y, Xiao F, Wang S (2013) An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network. Energy Build 57:278–288. https://doi.org/10.1016/j.enbuild.2012.11.007

Zhivich M, Cunningham RK (2009) The real cost of software errors. IEEE Secur Privacy 7(2):87–90. https://doi.org/10.1109/MSP.2009.56

Zhou Y, Zhu C, Tang L, Zhang W, Wang P (2018) Cyber security inference based on a two-level Bayesian network framework. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com