A Causal Explanatory Model of Bayesian-belief Networks for Analysing the Risks of Opening Data

Luthfi, Ahmad; Janssen, Marijn; Crompvoets, Joep

DOI
10.1007/978-3-319-94214-8_20

Publication date
2018

Document Version
Final published version

Published in
Proceedings of Business Modeling and Software Design - 8th International Symposium, BMSD 2018

A Causal Explanatory Model of Bayesian-belief Networks for Analysing the Risks of Opening Data

Ahmad Luthfi[1,2(✉)], Marijn Janssen[1], and Joep Crompvoets[3]

[1] Faculty of Technology, Policy and Management,
Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands
{a.luthfi, m.f.w.h.a.janssen}@tudelft.nl
[2] Universitas Islam Indonesia, Yogyakarta, Indonesia
ahmad.luthfi@uii.ac.id
[3] Katholieke Universiteit Leuven, Leuven, Belgium
joep.crompvoets@kuleuven.be

Abstract. Open government data initiatives result in the expectation of having open data available. Nevertheless, some potential risks like sensitivity, privacy, ownership, misinterpretation, and misuse of the data result in the reluctance of governments to open their data. At this moment, there is no comprehensive overview nor a model to understand the mechanisms resulting in risk when opening data. This study is aimed at developing a Bayesian-belief Networks (BbN) model to analyse the causal mechanism resulting in risks when opening data. An explanatory approach based on the four main steps is followed to develop a BbN. The model presents a better understanding of the causal relationship between data and risks and can help governments and other stakeholders in their decision to open data. We use the literature review base to quantify the probability of risk variables to give an illustration in the interrogating process. For the further study, we recommend using expert's judgment for quantifying the probability of the risk variables in opening data.

Keywords: Bayesian-belief Networks · Causality · Open data Relationship · Explanatory model · Risks · Sensitivity · Privacy Ownership · Misinterpretation · Misuse

1 Introduction

The expectations to gain access to public data have increased extensively over the last few years [1]. The creation of transparency, accountability, citizen engagement, and to enable business innovation are the main drivers for governments to open more their data [2–4]. The public expects governments to release their dataset for reaping the many benefits of opening data. For example, parents can explore open datasets from the government's portal to find the quality of educational institutes for their children. Foreigners can access open tourism datasets to decide some alternative destinations for their vacations.

Although initiatives to open data can create many benefits, they might also create disadvantages [1, 5]. Risks refer to the chance that these disadvantages come true and

B. Shishkov (Ed.): BMSD 2018, LNBIP 319, pp. 289–297, 2018.
https://doi.org/10.1007/978-3-319-94214-8_20

its impact. Potential risks include inaccuracy, sensitivity, privacy, inconsistency, and misuse of data. These risks result in the reluctance of governments to open their data [1]. In addition, two other reasons why governments and data providers are tend not to open their data: (1) opening public and/or private data are a comprehensive insight that may also able to meet risks like inappropriate interpretation of the data [4], and (2) a mistake in translating data or misuse of the data can endanger the reputation of data providers [6]. Governments sometimes also need a huge effort to investigate and analyze the cause and effect in terms of the opening data. Cause refers to an event or action of the risk in opening data that induces something else to occur. Effect means an event or action in releasing a dataset that occurred as a consequence of another event or action. For example, because of the inappropriate visualization of a dataset in government's portal information (as a cause), then the public will tend to misinterpretation of the dataset released (as an effect). Unfortunately, at this moment, the investigation of the risk in opening data and to what disadvantages the opening of data might result is not well-understood yet.

Bayesian-belief Network (BbN) is a method presented in this study because it allows constructing a causality relationship model between risk and effect variables [1, 7]. The use of a BbN in the risks analysis can provide explanations how the disadvantages of the opening of data are created. The BbN is able to capture the strength of causal links to define the cause under uncertainty variables [8, 9], and to visualize the possible consequences of the risks [10]. The consequences of the risks are the disadvantages and the impact of our situation. In this study, we develop an explanatory model as an approach to understand the relationship and influence factors of the risks in opening data. The explanatory processes follow four main steps [11]. To begin with, we require defining the risk variables and its relationships. Next, a network structure of the risks is developed showing how the variables interrelated. In the third step, we interrogate the model to obtain the understand the sensitivity of variables, and finally, the relationship diagrams are developed to enable to communicate the outcomes to stakeholders. The model is illustrated using health patient records dataset to analyze and to understand the relationship of causality factors of the risks in opening data.

The objective of the research presented in this paper is to develop a Bayesian-belief Network model for understanding the causal mechanisms of the risks in opening data. The purpose of the explanatory model is to enhance the understanding of governments and data providers of the potential risks in opening data and why these risks occur. This paper is systematically composed of five sections. In Sect. 1, the research drivers were described, Sect. 2 contains related works for BbN in risk assessment. In Sect. 3 the research approach taken in this study is explained. In Sect. 4, the development of a BbN is step-by-step presented, and finally, conclusions are drawn in Sect. 5.

2 Related Work

The BbN model was introduced for the first time in the field of risk management for terrorism threats [12]. In the second area, a BbN theory was implemented in a research related to the determination and fire mitigating actions [13]. Literature was reviewed to investigate the use of BbN for risk assessment. Table 1 shows seven related domains in

which BbN was used to analyze the potential risks. Only one paper is related to the opening data (see paper number 4). They proposed a model to weigh the potential risks and benefits of opening data. However, this paper did not focus on how to construct a BbN causality model to analyze the risks of opening data. The scientific contribution of this paper originates from developing a conceptual model of causality between risk variables in opening data.

Table 1. Related works in risks analysis using Bayesian-belief Networks

	Domain of the study	Description	Source
1	Bank	A case study to develop valid statistical models to measure, and predict such operational risks using Bayesian networks in the basel banking	[14]
2	Health	A presentation of health risks in the area of non-carcinogenic substances for non-critical organs and the additive assumption for multiple hazardous substances affecting the same organ	[15]
3	Smartphones	A discussed related to the information security and risk assessment of smartphone use in Finland	[10]
4	Open data	An iterative process of decision support model to weigh the potential risks and benefits of opening data	[1]

3 Research Approach

In this study, we will develop an explanatory BbN for the opening data. We adopted four main steps to develop a causality and relationship of the risk variables [11], as visualized in Fig. 1.

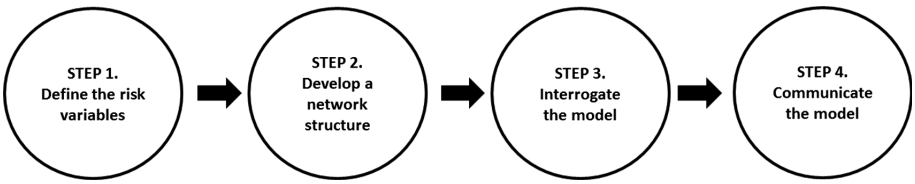

Fig. 1. Step-by-step of the Bayesian-belief Network development model

- Step 1. **Define the risk variables**. First of all, we conducted the literature review to identify the possible types of risk related to the opening data. Then, we classified the risk into several categories to make a singular variable of the risk for classifying the cause and effect elements. To make clear understanding and to avoid misinterpretation of the risk definition, we described the risks elements.

- Step 2. **Develop a network structure**. In this stage, a BbN structure is developed to present a set of risks elements that influence the potential disadvantages of releasing a dataset and provide its relationships. The parent node is a single element of the risk variable that influenced by a single or some causes variable. There are three sub-steps were followed to identify sub-nodes and their relationship. Sub-nodes and its relationships are possible to be repeated several times during the constructing a BbN.
  - Sub-step 2.a: identify the key elements. The key elements will become parent nodes of the top level node. The objective of this sub-step is to construct further sub-elements of the risk variables influencing them.
  - Sub-step 2.b: identify the remaining elements. This sub-step is to describe the causality of the various risk elements until the lowest of levels is generated. For example, Data Misuse (Z) is a parent node that influenced by two sub-nodes namely Data Sensitivity (X) and Data Ownership (Y).
  - Sub-step 2.c: identify the relationship. The objective of this sub-step is to identify the various nodes includes key elements, remaining elements based on their influences on each other. The relationship knowledge, in this work, will be identified from the literature based.
- Step 3. **Interrogate the model**. This step is aimed to interrogate the sensitivity and influence of variables on the risks. Sensitivity means that the responsiveness of each node or variable in the network structure is analyzed using a systematic approach to express the trigger variables [10, 11]. While the influence tends to analyze the frequency of impacts of the parent nodes on their respective child nodes by identifying their influential elements. The expected result of this step is to provide a better understanding of the most significant elements of the risks in opening data.
- Step 4. **Communicate the model**. The final process of the BbN development in this study is how to communicate the resulting network model. In this paper, we utilize the relationships diagram to describe the results visually.

4 An Illustration: Development of a Bayesian-belief Networks

In this section, we give an illustration to show the relationships of influence risks variable in opening data. The four steps explained in the research approach are followed. We illustrated using a Health Patient Record dataset to analyze the possibility of risks. The dataset, for instance, consists of some attributes like name_of_patient, date_of_birth, address, and phone_number. The scenario is government want to analyze the potential risk of the attributes before it released. By constructing a BbN, the government is able to understand the causality and its relationships of the risk element.

4.1 Define the Risk Variables

Open data have been shown to contribute to society through several programs of the governments of many countries in recent years [16]. Nevertheless, along with the benefits

Table 2. Type of risks in opening data

	Category	Description	Source
1	Data inaccuracy	Data inaccuracy can occur when the data providers release their data. Some of the causes of data inaccuracy include: (a) data entry mistake by the users or data operators, (b) flawed data entry process, (c) the null problem with the value of the data, and (d) deliberate error when the users enter a wrong value of the data. This category can affect the quality of the data	[18, 20, 21]
2	Data misuse	Data disclosure can make personal or individual data identifiable by combining several datasets. Some cause misuses of the data are: (a) discredit personal profile, (b) access as unauthorized users, and (c) diminish the government's or company's reputation. This situation influencing the data privacy	[18, 22–24]
3	Data sensitivity	Releasing data can include sensitive attributes. Personal identity elements like full name, date of birth, address, and phone number are possible to be analyzed by the users. This category, influencing the data privacy and data violation	[18, 25]
4	Data incompleteness	Opening incomplete data, create misunderstanding about the meaning of the data. The caused elements of this category are (a) the anonymity of the data source, (b) inappropriate aliases formula, and (c) mismatch of the attribute relationships. This situation is also possible influencing the data quality and data misinterpretation	[18, 22, 23, 25, 26]
5	Data misinterpretation	Publishing data by governments or companies are possible to drive a misinterpretation of the data. The causes factors of this category are: (a) insufficient domain expertise, (b) important variables are omitted, (c) inappropriate data visualization, and (d) error of attribute correlation. The effect of this risk category is influencing the data quality and data incompleteness	[17–19, 23, 27]

of implementing the disclosure of data, potential risks of opening data are emphasized [17–19]. The risks are classified into five categories, as presented in Table 2.

4.2 Develop a Network Structure

In this step, we begin to develop a BbN structure to identify the causes and relationships between risk elements. In step 1, define the risk variables, we have identified five main categories of the risk in opening data. Based on the cause-and-effect for each category, a BbN structure is created. To do so, there are three sub-steps to make a detail action of this process, as follows: (1) Identify the key elements. From the identified of the risk elements, we need to classify which are the parent nodes in particular,

(2) Identify the remaining elements. After the parent and the child nodes have identified, we then make a connection between the parent and child as the one to one or one to many connectivity, and (3) Identify the relationships. The main objective of this process is to make a relationship between parent nodes to other parent nodes become a network a relationship. This step can be processed in several iterations until the lowest sub-node is identified and correlated.

- Sub-step 2.a. **Identify the key elements**. We identified parent nodes of the top level nodes for each of the five categories.
  - Data inaccuracy parent node. Factors like a data entry mistake, flawed data entry process, the null problem, and deliberate error as the influencing factors are sub-elements.
  - Data misuse parent node. The cause elements of this category are discredited personal profile, unauthorized user, and diminish reputation.
  - Data sensitivity parent node having personally identifiable, as a single sub-element.
  - Data incompleteness parent node. The anonymity of data source, inappropriate aliases formula, and the mismatch of attribute relationship are the influencing factors.
  - Data misinterpretation parent node. The insufficient of domain expertise, omitted important variables, inappropriate data visualization, and error of attribute correlation are all sub-elements of data misinterpretation.
- Sub-step 2.b. **Identify the remaining elements**. From the parent node and sub-elements constructed in Sub-step 2a, we made the connection between the variables. The correlation of each node and sub-elements show the relationship of the risks elements until the lowest of levels.

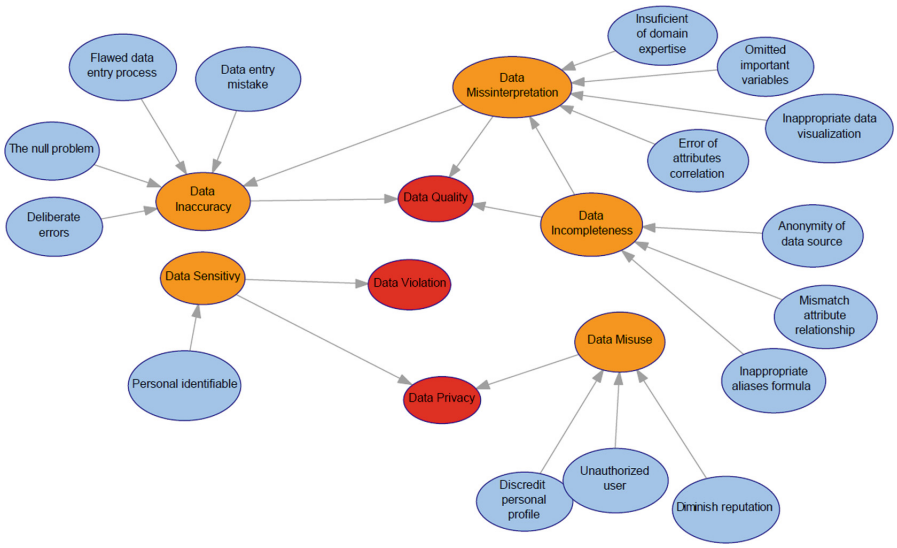

Fig. 2. Causality model of the risks in opening data

- Sub-step 2.c. **Identify the relationships**. In this step, the effect elements of the parent nodes are constructed. There are three attributes influenced by the parent nodes. First, data quality is affected by the data inaccuracy, data misinterpretation, and data incompleteness. Second, data privacy is influenced by the data misuse, and data sensitivity. Finally, single data violation is affected by the data sensitivity. The causal network structure and its relationship is visualized in Fig. 2.

4.3 Interrogate the Model

After the BbN structure is developed, we interrogated the resulting model by distributing the probabilities for each node and sub-elements. The objective of this step is to interrogate the sensitivity level the risk elements and to present the highest and the lowest probabilities of the risks. We designed using the literature review base to quantify the value of each risk element. We divided the value of the probabilities into three options: Yes, Neutral, and No.

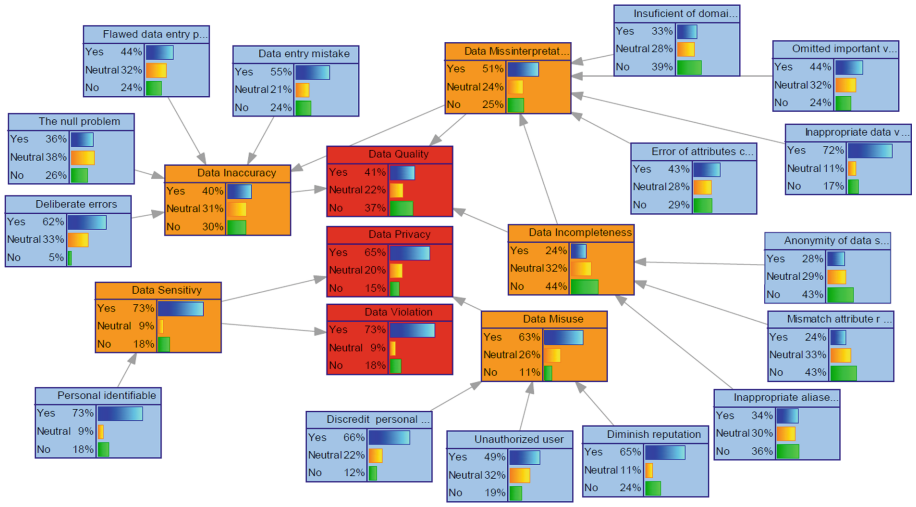

Fig. 3. Causality of the risk elements relationships in opening data

Figure 3 presents the causality and relationships between risk variables in opening data. Based on the literature review, we quantified both cause and effect nodes. To illustrate, we gave values for Data Misuse parent node (Yes = 0.63; Neutral = 0.26, No = 0.11). For cause variables, we identified that the value of sub-nodes as follows: (1) Discredited personal profile (Yes = 0.66; Neutral = 0.22, No = 0.12), (2) Unauthorized user (Yes = 0.49; Neutral = 0.32, No = 0.19), and (3) Diminish reputation (Yes = 0.65; Neutral = 0.11; No = 0.24). Next, we require quantifying the Data Privacy variable as the parent node of Data Misuse with (Yes = 0.65; Neutral = 0.20; No = 0.15). In this case, we can assume that some attributes of health patient records as a given dataset in this illustration has 0.65 or 65% the potential risks when it decided to publish.

4.4 Communicate the Model

The final step of the BbN model is to create a model that can be communicated to stakeholders like decision-makers, policy-makers, and data providers. In further research, we will develop a comprehensive explanation of the model's development in line with the various types of analysis formats [10, 11]. There are some approaches to disseminate the data to the public. For example, data visualization platforms like graphs, charts, histogram, or scatter plot are able to help the stakeholders to use the models in practice [11].

5 Conclusion

The public expects governments and data providers to open their data for reaping the benefits. However, behind the merits of releasing the data, the governments are often reluctant to open their data due to possible disadvantages. Disadvantages like low data quality, individual identification, and opening inaccurate data are the main drivers of government for not opening their data. The risk elements are used for determining the exposure to these disadvantages. At the same time, the causality and relationships between the potential of the risk elements are not investigated yet. In this paper, we derive a Bayesian-belief Network (BbN) to construct a causal model of the risk in opening data. The explanatory model consists of four main steps. First, define the risk variables and its relationship. Second, a network structure of the risk is constructed. In the third step, we interrogate the model to obtain the current sensitive analysis, and finally, we require to communicate the model using relationship diagram to provide the new knowledge to the stakeholders in terms of the risks analysis in opening data. The main purpose of the model is to present the better understanding of the governments and other stakeholders in decision analysis. For example, from the illustration presented in this paper, we noticed that governments need to consider the data privacy issues including its causes and relationships. The limitation of this study is using the literature review base to quantify the risk variables to give an illustration in the interrogating step. For the further research, we recommend using expert's judgment for quantifying the probability of the risks variables in opening data.

References

1. Luthfi, A., Janssen, M.: A conceptual model of decision-making support for opening data. In: Katsikas, Sokratis K., Zorkadis, V. (eds.) e-Democracy 2017. CCIS, vol. 792, pp. 95–105. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71117-1_7
2. Ali-Eldin, A.M.T., Zuiderwijk, A., Janssen, M.: Opening more data: a new privacy scoring model of open data. In: Seventh International Symposium on Business Modelling and Software Design (BMSD 2017), Barcelona, Spain. SCITEPRESS - Science and Technology Publication, Lda (2017)
3. Lourenço, R.P.: An analysis of open government portals: a perspective of transparency for accountability. Gov. Inf. Q. **32**(3), 323–332 (2015)
4. Zuiderwijk, A., Janssen, M.: Open data policies, their implementation and impact: a framework for comparison. Gov. Inf. Q. **31**(1), 17–29 (2013)
5. Zuiderwijk, A., Janssen, M.: Towards decision support for disclosing data: closed or open data? Inf. Polity **20**(2–3), 103–107 (2015)

6. Barry, E., Bannister, F.: Barriers to open data release: a view from the top. Inf. Polity **19**(1–2), 129–152 (2014)
7. Fenton, N., Neil, M.: Risks Assessment and Decision Analysis with Bayesian Networks. CRC Press, Noca Raton (2012)
8. Nadkarni, S., Shenoy, P.P.: A causal mapping approach to constructing Bayesian networks. Decis. Support Syst. **38**(2), 259–281 (2004)
9. Hu, H., Elrafey, A.H., Kerschberg, L.: Using modular ontologies to capture causal knowledge contained in Bayesian networks. In: 2017 IEEE/ACM International Conference. ACM, Sydney (2017)
10. Herland, K., Hämmäinen, H., Kekolahti, P.: Information security risks assessment of smartphones using Bayesian networks. J. Cyber Secur. **4**, 65–85 (2016)
11. Chakraborty, S., et al.: A Bayesian network-based customer satisfaction model: a tool for management decisions in railway transport. J. Decis. Anal. **3**(4), 2–24 (2016)
12. Hudson, L.D., et al.: An Application of Bayesian Networks to Antiterrorism Risk Management for Military Planners (2002)
13. Gulvanessian, H., Holický, M.: Determination of actions due to fire: recent developments in Bayesian risk assessment of structures under fire. Prog. Struct. Eng. Mater. **3**(4), 346–352 (2001)
14. Cornalba, C., Giudici, P.: Statistical models for operational risk management. Phys. A Stat. Mech. Appl. **388**(1–2), 166–172 (2004)
15. Liu, K., et al.: Applying Bayesian belief networks to health risks assessment. Stochast. Environ. Res. Risk Assess. **26**(3), 451–465 (2012)
16. Zuiderwijk, A., Janssen, M., David, C.: Innovation with open data: essential elements of open data ecosystems. Inf. Polity **19**(2–3), 17–33 (2014)
17. Barnickel, N., et al.: Berlin open data strategy: organisational, legal and technical aspects of open data in Berlin. In: Concept, Pilot System and Recommendations for Action (2012)
18. Martin, S., et al.: Risk analysis to overcome barriers to open data. Electron. J. e-Gov. **11**(1), 348–359 (2013)
19. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012)
20. Chen, D., Zhao, H.: Data security and privacy protection issues in cloud computing. In: International Conference and Privacy Protection Issues in Cloud Computing, pp. 647–651. IEEE Computer Society, Hangzhou (2012)
21. Dekkers, M., et al.: Data and Metadata Licensing, O.D. Support, Editor. European Comission (2014)
22. Walter, S.: Heterogeneous database integration in biomedicine. J. Biomed. Inf. **34**(4), 285–298 (2001)
23. Amit, S.P., Larson, J.A.: Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput. Surv. **22**(3), 183–236 (1990)
24. Yannoukakou, A., Araka, I.: Access to government information: right to information and open government data synergy. In: 3rd International Conference on Integrated Information (IC-ININFO), vol. 147, pp. 332–340 (2014)
25. Tran, E., Scholtes, G.: Open Data Literature Review. Barkeley School of Law, University of California (2015)
26. Okediji, R.L.: Government as owner of intellectual property? Considerations for public welfare in the era of big data. Vanderbilt J. Entertain. Technol. Law **18**(8), 331 (2014)
27. Uhlir, P.F.: The Socioeconomic Effects of Public Sector Information on Digital Networks: Toward a Better Understanding of Different Access and Reuse Policies. National Research Council, Washington, DC (2009)