Construction and Verification of a Bayesian Network for Third-Party Excavation Risk Assessment (BaNTERA)

Colin A. Schell^a, Andres Ruiz-Tagle^b, Austin D. Lewis^c, and Katrina M. Groth^d

^a University of Maryland, College Park, United States, cschell@umd.edu ^b University of Maryland, College Park, United States, aruiztag@umd.edu ^c University of Maryland, College Park, United States, adlewis@umd.edu ^d University of Maryland, College Park, United States, kgroth@umd.edu

Abstract: According to the Pipeline and Hazardous Material Safety Administration (PHMSA), thirdparty damage is a leading cause of natural gas pipeline accidents. Although the risk of third-party damage has been widely studied in the literature, current models do not capture a sufficiently comprehensive set of up-to-date root cause factors and their dependencies. This limits their ability to achieve an accurate risk assessment that can be traced to meaningful elements of an excavation. This paper presents the construction, verification, and validation of a probabilistic Bayesian network model for third-party excavation risk assessment, BaNTERA. The model was constructed and its performance verified using the best available industry data and previous models from multiple sources. Historical industry data and nationwide statistics were compared with BaNTERA's damage rate predictions to validate the model. The result of this work is a comprehensive risk model for the third-party damage problem in natural gas pipelines.

1. INTRODUCTION

Between 2016 and 2020, third-party damage was responsible for an average of 21% of all pipeline failures per the Pipeline and Hazardous Safety Administration (PHMSA) [1, 2]. The resulting monetary and human losses have motivated a number of research projects and studies into the causes of third-party damage. However, the context of an excavation scenario varies widely and industry preventative measures are constantly evolving making third-party damage a difficult problem to model. As a result, there is a lack of models that include an up-to-date and comprehensive set of causal factors for modelling the probability of a third-party excavation causing a pipeline failure. This paper presents a Bayesian Network for Third-Party Excavation Risk Assessment (BaNTERA), a probabilistic model for assessing the risk that third-party excavation poses to natural gas pipelines. BaNTERA utilizes a wide variety of sources to inform an up-to-date and comprehensive set of causal factors and can thus provide insight into the excavation process assisting decision makers in mitigating potential pipeline risks.

The paper is structured as follows. The remainder of Section 1 discusses the basics of a third-party excavation process and the current state-of-the-art in modelling a third-party excavation. Section 2 includes an introduction to Bayesian networks, a discussion of the hierarchical structure that organizes context factors within BaNTERA, and the data processing and model parameterization strategies. Section 3 presents the resulting BaNTERA structure, the prior node probabilities, and the process used to verify the model. Section 4 further discusses the results of Section 3 and also includes the challenges and opportunities for BaNTERA. Lastly, Section 5 presents the conclusions of this paper.

1.1 Third-Party Excavation Process

Defining and modeling an excavation process is a difficult task as no two excavations are the same. The Common Ground Alliance (CGA) was used as a reference for identifying modern damage prevention techniques and best dig-in practices to include in BaNTERA. CGA is an organization that encourages pipeline damage prevention through the use of an 811 One Call system [3]. At least two days prior to

digging, excavators are encouraged to call their local 811 Once Call notification center or submit an online ticket to notify the center of the upcoming excavation. A professional then checks the excavation area for any at-risk utilities and marks their location for the excavator's reference. After marking, the excavator is permitted to begin digging by safely exposing the underground utilities using hand tools in order to avoid any damage. This process is illustrated in Figure 1 and if not performed correctly, pipeline damage may ensue. If an underground utility is struck with enough force, third-party damage in the form of puncture failure can occur.

The excavation process consists of several actors and objects. The actors in an excavation process are defined as the third-party excavators, the owner or manager of the pipeline, and the One Call notification center. Each of these actors directly interact with the excavation process and a failure of any one of these actors to perform their duties may increase the likelihood of a pipeline failure. Public and regulatory enforcement also interact with the excavation process but in indirect ways. For this reason, they were not included as a main actor in BaNTERA. The objects considered in an excavation process are facility maps/records and the notification ticket. All of these actors and objects interact with each other within the excavation environment which contains its own set of important factors including, geographical, regulatory, and site-specific contexts.



Figure 1: Five Steps for Safe Digging According to the Common Ground Alliance

1.2 Previous Third-Party Damage Models

In an effort to identify third-party damage root-causes and the associated context factors, a variety of modeling approaches have been used in the literature. Chen and Nessim [4] developed a third-party damage model using fault trees which served as the basis for future modeling efforts. These future efforts expanded upon the list of modeled damage factors and safety barriers allowing for more accurate assessments of the probability of a third-party damage. However, fault tree models struggle with incomplete datasets (a common characteristic of third-party data) and are limited in their ability to represent complex dependencies given their basis in Boolean logic. To overcome this, researchers began mapping fault tree models to a Bayesian network structure. For instance, Xiang and Zhou [5] used an expected maximization algorithm to learn fault tree-based Bayesian network model parameters from incomplete datasets in order to predict third-party damage probability. Wang et al [6] leveraged the data fusion capabilities of Bayesian networks to overcome data limitations surrounding urban buried gas pipelines. Historical data was combined with expert knowledge to estimate fault tree failure probabilities which were then mapped to a Bayesian network structure. Fault tree-based Bayesian networks do address the problem of dealing with incomplete datasets but because their structure is built using a fault tree model, these models are still limited by their foundation in Boolean logic.

In order to accurately model the complex dependency relationships inherent in an excavation process, a fully Bayesian network approach has been recently employed by researchers. Jackson and Mosleh [7] use a fully Bayesian network approach to model both the physical and human-related variables in an excavation process but this model is still in its development phase. Guo et al [8] developed a Bayesian

network model for the evolution of pipeline damage risk which focuses more on the consequences of the pipeline failure than the root-cause factors. Additionally, Cui et al [9] modeled unintentional third-party damage using "3rd Party Disturbance Index," a function based on the Bayesian network node values.

The collection of these past works forms the foundation on which to expand the scope of third-party damage risk modeling. Additionally, given the evolution of damage prevention techniques, a comprehensive and up-to-date set of damage factors must be developed in order to accurately represent a modern excavation process. By leveraging modern datasets and by building upon the accomplishments of past modeling efforts, a decision support tool can be created to inform decision makers on how to reduce the risk of third-party damage.

2. METHODS

2.1. Bayesian Networks

The third-party damage model is structured as a Bayesian network. These types of networks are probabilistic graphical models that represent causal relationships within a system. As such, they can be used to model third-party damage by representing how risk influencing and other causal factors interact at an excavation site and lead towards damage.

Bayesian networks are widely used as causal models in risk and safety applications to express the joint probability distribution of a system's variables. Variables and their causal conditional dependencies are represented, respectively, by the nodes $V = \{V_1, V_2, ..., V_n\}$ and arcs of a directed acyclic graph (DAG), also referred as a Bayesian network's structure. The dependency strength between two variables is quantified through a conditional probability table (CPT) or function (CPF) depending on whether the variables involved are discrete or continuous. The Bayesian network model corresponds to the prior joint probability distribution $Pr_0(V_1, V_2, ..., V_n)$. Mathematically, this distribution can by computed using the factorization formula in Equation (1):

$$\Pr_0(V_{1,\ldots},V_n) = \prod_{i=1}^n \Pr(V_i | pa(V_i))$$
(1)

where $pa(V_i)$ corresponds to the parent nodes of V_i ; that is, the set of nodes in the Bayesian network that have an outgoing edge directed at the node V_i .

As Bayesian network structures are based on the conditional relationships between their nodes, joint probability distributions of system variables can be updated with new knowledge about the system via evidence on the nodes. Assigning observed evidence to a set of nodes $V_k = v_k$ leads to a Bayesian update from which the posterior distribution $Pr_1(V_1, ..., V_k = V_n)$ is obtained. In doing so, analysts can use Bayesian networks to perform probabilistic inference.

2.2. Taxonomy and Hierarchical Structure of a Third-Party Damage Bayesian Network Model

From the excavation processes described in the introduction, a detailed hierarchical taxonomy of thirdparty excavation scenarios was built. This taxonomy consists of three categories of attributes at a conceptual excavation site. Terms and concepts in the "Actor" category are associated with the thirdparty excavator, the facility owner and manager, and the One Call notification center. Equipment and physical items found at the excavation site are part of the "Objects" category. The final category, "Environment," provides further details about the context and background in which the excavation occurs.

The third-party excavation taxonomy illustrates a wide range of context factors that were considered when creating BaNTERA. Figure 2 is a conceptual diagram of relevant factors, hierarchical structure,

and associated dependencies used to model third-party damage within BaNTERA. This diagram is based on expert knowledge and causal factors from PHMSA and CGA incident reports. At the bottom of BaNTERA are summary target variables that include the sufficiency of pipe hit preventive measures, the pipe hit itself, and subsequent damage. Target variables depend on the other node types existing in the excavation process space; one such type consists of the groups of root causes that contribute to third-party damage. These groups capture accumulated relationships of similar pipeline damage root causes. These groups are similar to those expressed by the CGA's DIRT [3], which are based on distinct steps of the excavation process (that is, notifying, locating, and excavating).

Figure 2: Conceptual Diagram of Third-Party Damage Model Outlining the Causal Flow from Excavation Context Nodes and Risk-Influencing Factors Context Nodes to Root Causes and Target Variables.



The next level of model nodes to consider are the root causes and the context variables that affect them. This part of BaNTERA has the greatest number of variables that can vary during the excavation process, making this part of the model the most interactive. Root causes defined by the CGA DIRT are considered for its construction because of their wide use by utility companies. These root causes are mainly focused on the excavation process itself and not on its context. To expand the range of acknowledged factors that can contribute to third-party damage target variables and, in particular, its root causes, a set of risk-influencing factors (RIFs) were defined based on the context of an excavation. RIFs were distinguished in two categories by what kind of measures could be adopted to mitigate them: actor-based RIFs and object-based RIFs. These factors are themselves affected by a third RIF, the environment in which the excavation takes place. This environment can be physical, legal, or related to the excavation activities themselves.

2.3. Data Processing and Model Parameterization

In order to parameterize BaNTERA, four distinct data sources were used. These are described as follows:

- 1. *Utility partner incident data*: more than 7,000 third-party damage reports were provided. These reports were obtained between 2016 and 2020, and provide information on more than forty different fields, such as the root cause of an incident and the type of excavator involved in it.
- 2. *PHMSA gas transmission and distribution accident and incident data*: U.S. federal regulation requires the submission of incident reports for all pipeline incidents that comply with the codes CFR parts 191 and 195, and PHMSA maintains a record of these data [1], providing relevant information on more than 200 incidents from 2016 to 2020, such as the material of the damaged pipeline and the class location in which the incident occurred.

- 3. *CGA DIRT database*: the CGA annual DIRT database contains summary statistics on multiple relevant features from historical excavation damages to underground infrastructure. Regarding natural gas pipelines, this dataset provides unique information that cannot be found in the data sources above, such as an excavator's awareness of notification practices and the existence of previous damage or deterioration on a pipeline involved in an incident. Summary statistics from 2018 and 2019 were used in this work.
- 4. *GTI Energy models and statistics*: GTI Energy provided three different Bayesian network models for informing the construction and parameterization of BaNTERA [10, 11, 12]. Each model had a different outcome of interest, which were "locating and marking practices errors," "third-party damage prevention practices errors," and "puncture physics of failure," which were built on data from multiple utility partners between 2016 and 2018. Summary statistics on third-party damage root causes and puncture failure models were highly relevant to the construction of BaNTERA.



Then, as shown in the process of Figure 3, the data sources, taxonomy, and hierarchical structure presented above were processed in order to build two datasets, an observational and expert-based dataset. These are described as follows:

- *Observational dataset*: this dataset consists of historical third-party-caused incident statistics that were processed into a format consistent with the conditional probability tables present in BaNTERA's structure. For instance, the utility partner incident data source showed that 23.04% of excavators involved in an incident were non-professionals and that, as shown in the CGA DIRT database, only 51.72% of them notified their intent to excavate.
- *Expert-based dataset*: this dataset consists of subject-matter expert knowledge on different variables on BaNTERA, including physics of failure models and expert-elicited correlations between variables. For instance, Brooker's puncture model [13] was included in the expert-based dataset to inform the probability of a puncture in an excavation. This model assumes that:

$$Q_p = 7.0074 \times 10^{-4} t (\sigma_u + 410.4) (L + 22.41) \frac{W}{(3.142 + W)}$$
(2)

In non-plastic materials. Here, Q_p is the pipe material puncture resistance, t [mm] is the pipe wall thickness, $\sigma_u [MPa]$ is the ultimate strength of the pipe material, and W [mm] and L [mm] are the excavation tool tooth width and length, respectively.

These two datasets were processed into a format consistent with BaNTERA's model structure, enabling its direct parameterization. It is important to note, however, that multiple data sources are being fused into a single model. As such, relevant statistics present in the data sources used to parameterize BaNTERA are prone to change if the model is not well defined. Given this, it is highly important to verify that BaNTERA reflects the damage rates present in the data sources used in its parameterization.

The resulting BaNTERA structure, prior node probabilities, and model verification results are shown in Section 3.

3. RESULTS

3.1. BaNTERA Structure

BaNTERA was structured using the node relationships and dependencies identified in Figure 2. The result is a comprehensive model that captures not only the root causes of third-party damage, but also provides a method of expressing failure through context described within the taxonomy and the overarching RIFs. Figure 4 shows BaNTERA's structure.

Figure 4: BaNTERA's Structure. Visualization Made in GeNIe. Color Legend: Gray - Target Summary Nodes; Red - DIRT Root Cause Group; Pink - DIRT Root Cause Subgroup; White -DIRT Root Cause; Light Blue - Non-DIRT Root Cause; Blue - Third-Party Excavation Actor Context; Yellow - Environment Context; Purple - Facility Context; Green - Excavation Context.



BaNTERA's structure consists of nodes that provide different functions within the model and different types of information. At the bottom of the structure are the summary target nodes that combine root cause information with excavation-related context. These provide information into potential occurrences of pipe contact, puncture, and damage during a third-party excavation. The parents of the target nodes are the large root cause sub-group and group nodes, which capture accumulated relationships of similar root causes. These groups are similar to those expressed in the DIRT report, which are based on distinct steps of the excavation process (i.e., notifying, locating, excavating). The next step up in the structure contains specific root causes and the risk-influencing nodes that affect them. This part of the structure has the greatest number of nodes that can vary during the excavation process, making it highly interactive. Lastly, the top of the structure consists of overarching context nodes, representing the environment in which the excavation is taking place.

BaNTERA's structure captures different causal contexts and root causes for third-party damage. The color of the nodes reflects what kind of information they provide. As shown in Figure 4's legend, white, light blue, red, pink, and gray nodes describe specific and general root causes for third-party damage during an excavation. White and light blue nodes are individually described root causes of damage, while red nodes indicate the formation of a root group based on a number of similar root causes. Pink nodes provide an initial amalgamation of root causes into smaller root cause groups, which are then joined to form a root cause group. Gray nodes describe the summary target nodes that BaNTERA is

trying to represent. The other nodes represented in BaNTERA's structure describe the context in which the third-party damage occurs. Yellow nodes describe the overall excavation environment nodes that are specific to the context of the excavation site, while the blue, green, and purple nodes relate to context factors associated with the specific excavation site scenario. Blue nodes describe the context associated with the third-party excavation actors, which includes both the excavator itself and the utility locator. Green nodes describe the context of the excavation site, while purple nodes are used to represent pipeline characteristics. These causal context factors can also be placed within the overarching RIFs described earlier.

3.2. BaNTERA's Prior Nodes' Probabilities and Verification Results

BaNTERA's parameterization resulted in a mean joint probability distribution for third-party damage risk. BaNTERA's parameterization result can be found in Figure 4 in which there is a combination of both discrete and continuous nodes. Here, discrete nodes result in the expected probabilities of the different states it can take, whereas continuous nodes result in a histogram of 10,000 simulated values based on logic sampling [14]. All of these calculations were made using BayesFusion's PySMILE wrapper and visualized through the software GeNIe [15].

The obtained prior expected probabilities for BaNTERA's summary target nodes can be found in Table 1.

Summary target node	Node states	Prior expected probability
T1: Preventive Measures	Sufficient	0.8883
	Insufficient	0.1117
T2: Pipe Hit	No	0.9969
	Yes	0.0031
T3: Punc. Force >	No	0.9975
Resistance	Yes	0.0025
T4: Damage (Puncture)	No	0.9975
	Yes	0.0025

 Table 1: BaNTERA's Prior Expected Probabilities on Summary Target Nodes

BaNTERA's results presented in Table 1 show that 3.1 out of 1,000 excavations in the U.S. are expected to result in a pipe hit, and that 2.5 of those excavations will result in a punctured pipe.

In order to verify a successful parameterization of BaNTERA, it was necessary to compare its target summary statistics to the sources with which they were parameterized. In particular, BaNTERA's predicted third-party puncture damage rates per 1,000 notified excavations were compared with the utility partner incident damage rate per 1,000 notified excavations and with GTI Energy statistics (see Section 2.3). This comparison is shown in Table 2.

Table 2: BaNTERA's Verification with Respect to the Utility Partner's and GTI Energy'	S
Third-Party Damage Rates per 1,000 Notified Excavations	

	Damage rate per 1,000 notified excavations
Utility partner incident data (average for 2017-2020)	1.41
GTI Energy (average for 2016-2018)	1.62
BaNTERA	1.53

4. DISCUSSION

4.1. BaNTERA Structure

BaNTERA's structure was created following a methodical examination of current practices and root causes outlined in the structure taxonomy; as such, it provides a clear and logical flow of excavation events. The Bayesian network structure is adjustable allowing for expanding, restricting, adding and deleting new nodes depending on the purpose and interests of the model users. From a structural perspective, this would just impact the conditional probabilities of the changed nodes and its children. This is a very useful tool for as dynamic an environment as excavation regulations and policies. It should be noted though that in this current version BaNTERA, puncture damage is the only damage type considered for modeling third-party damage.

4.2. Prior Node Probabilities and Model Verification

BaNTERA is composed of a number of heterogeneous data sources, which is expected for a complex engineering system such as the U.S. natural gas underground infrastructure. Further, these data had inconsistent fields and naming, which hindered the parameterization process of BaNTERA. For instance, PHMSA incident data only considers 20 root cause variables for third-party damage, whereas the CGA DIRT database considers 26. As such, the model predictive performance can become sensitive to its parameterization strategy and data processing practices.

To show that BaNTERA's predictions on third-party damage were not hindered by the issues described above, the model was verified by (1) asking both GTI Energy and the utility partner if the predictions were consistent with their knowledge on the system, and (2) by comparing BaNTERA's predicted damage rates to the damage rates present on the data sources used in its construction. These verification steps were done with BaNTERA's third-party damage rate prediction over 1,000 excavation activities, as is common in the U.S. pipeline industry. As shown in Table 1, BaNTERA predicted a prior of 3 pipe hits and 2.5 third-party puncture damages per 1,000 excavations. These predictions were consistent with both GTI Energy and the utility partner knowledge. Furthermore, to ensure the resulted prior probabilities were consistent with the damage rate statistics for GTI Energy and the utility partner, we compared their historical damage rates per 1,000 notified excavations to BaNTERA's. The results presented in Table 2 show BaNTERA's predictions had an 8.6% and 5.5% absolute percent points difference with the utility partner and GTI Energy damage rates, respectively. Therefore, BaNTERA's predictions have yet to be validated through comparison with nationwide statistics, which will be done in future work.

4.3. Challenges and Opportunities

BaNTERA is built as a proof-of-concept model for third-party risk management and thus, model performance can still be improved with future work. The first improvement opportunity is to balance the complexity of modeling an excavation scenario with the intended scope of the model. BaNTERA was designed to capture a wide variety of potential excavation scenarios and as a result, some network nodes are intentionally broad in scope. By delving deeper into these broad network nodes, researchers may gain key insight into how specific excavator actions or context details affect the rest of an excavation scenario.

The second improvement opportunity is related to the damage types modeled in BaNTERA. Currently, puncture damage is the only damage type modeled in BaNTERA. Other such damage types include dents and gouges but often result in latent failures making it more difficult to obtain data on dent/gouge root causes. Using puncture damage, BaNTERA can predict the lower bound of pipeline damage rates

but in order to narrow in on the actual damage rates, further work in predicting these other damage types is required.

The final improvement opportunity for BaNTERA is tied to the limited availability of success space data for third-party excavations. Success space data is defined as information about excavations that did not result in a damage. In general, utilities do not collect granular enough information on excavations that don't result in incidents. For example, many ticket datasets do not specify the responsible party for an excavation. In order to compensate for this lack of data, many third-party models, including BaNTERA, must rely on the use of expert knowledge to parametrize key nodes that cannot be parameterized using historical data. Although expert knowledge can provide valuable insight, the lack of success space data nevertheless restricts the depth of insight that BaNTERA can provide and future work by the industry and regulatory agencies in obtaining success space data will strengthen third-party damage modeling efforts.

5. CONCLUSION

This paper presents a probabilistic model for third-party excavation risk assessment, BaNTERA. By leveraging a variety of modern data sources, expert judgement, and insight from previous models, BaNTERA can assist decision-makers in gaining insight into the root-causes of third-party damage and the context factors that surround these incidents. As a result, more effective preventative measures can be identified to lower the risk of future third-party incidents. BaNTERA's verification results indicate a successful first step in creating a comprehensive model for third-party risk assessment. These results also highlight a number of promising improvement opportunities, specifically, adding specificity to key network nodes, expanding the number of modeled damage types to include latent failures, and collecting additional data specifically of excavations that do not result in failure. Working to expand the capabilities of BaNTERA may in turn expand the capabilities of third-party excavation management to provide more effective policies aimed at improving pipeline safety.

Acknowledgements

This information, data or work presented herein was funded in part by the U.S. Department of Transportation Pipeline and Hazardous Materials Safety Administration's Pipeline Safety Research and Development Program, under Award Number 693JK31910002POTA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Pipeline and Hazardous Materials Safety Administration, the Department of Transportation, or the U.S. Government.

References

- [1] PHMSA, "Pipeline incident 20 year trends." 2020. [Online]. Available: https://www.phmsa.dot.gov/data-and-statistics/pipeline/ pipeline-incident-20-year-trends
- [2] J. S. Santarelli, "Risk Analysis of Natural Gas Distribution Pipelines with Respect to Third Party Damage." Master's thesis, Western University, Ontario, Canada, (2019).
- [3] CGA, "DIRT Annual Report for 2019." Common Ground Alliance, (2020).
- [4] Q. Chen and M. Nessim, "Reliability-based prevention of mechanical damage to pipelines." C-FER Technologies, PR-244-9729, (1999).
- [5] W. Xiang and W. Zhou, "Bayesian network model for predicting probability of third-party damage to underground pipelines and learning model parameters from incomplete datasets," *Reliability Engineering & System Safety*, (2021).
- [6] W. Wang, K. Shen, B. Wang, C. Dong, F. Khan, and Q. Wang, "Failure probability analysis of the urban buried gas pipelines using Bayesian networks," *Process Safety and Environmental Protection*, vol. 111, (2017).

- [7] C. Jackson and A. Mosleh, "Oil and Gas Pipeline Third Party Damage (TPD) A New Way to Model External Hazard Failure," *Los Angeles*, (2018).
- [8] X. Guo, L. Zhang, W. Liang, and S. Haugen, "Risk identification of third-party damage on oil and gas pipelines through the Bayesian network," *Journal of Loss Prevention in the Process Industries*, vol. 54, (2018).
- [9] Y. Cui, N. Quddus, and C. V. Mashuga, "Bayesian network and game theory risk assessment model for third-party damage to oil and gas pipelines," *Process Safety and Environmental Protection*, vol. 134, (2020).
- [10] T. Peng, "Modernizing Tools for 3rd Party Damage." Technical Report 22412, GTI Energy, (2019).
- [11] B. Kolade, "Towards the Development of a Risk Model for Locates." Technical Report 22397, GTI Energy, (2019).
- [12] D. C. Brooker, "Numerical modelling of pipeline puncture under excavator loading. Part I. Development and validation of a finite element material failure model for puncture simulation," *International Journal of Pressure Vessels and Piping*, vol. 80, no. 10, (2003).
- [13] D. Brooker, "Numerical modelling of pipeline puncture under excavator loading. Part II: parametric study," *International Journal of Pressure Vessels and Piping*, (2003).
- [14] M. Henrion, "Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling," in *Machine Intelligence and Pattern Recognition*, vol. 5, J. F. Lemmer and L. N. Kanal, Eds. North-Holland, (1988).
- [15] BayesFusion, "GeNIe." 2019. [Online]. Available: https://bayesfusion.com/genie/