Explaining Bayesian Networks in Natural Language using Factor Arguments. Evaluation in the medical domain.

Jaime Sevilla[1], Nikolay Babakov[2,*], Ehud Reiter[1] and Alberto Bugarín[2]

[1]University of Aberdeen, Aberdeen, UK
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain

Abstract

In this paper, we propose a model for building natural language explanations for Bayesian Network Reasoning in terms of factor arguments, which are argumentation graphs of flowing evidence, relating the observed evidence to a target variable we want to learn about. We introduce the notion of factor argument independence to address the outstanding question of defining when arguments should be presented jointly or separately and present an algorithm that, starting from the evidence nodes and a target node, produces a list of all independent factor arguments ordered by their strength. Finally, we implemented a scheme to build natural language explanations of Bayesian Reasoning using this approach. Our proposal has been validated in the medical domain through a human-driven evaluation study where we compare the Bayesian Network Reasoning explanations obtained using factor arguments with an alternative explanation method. Evaluation results indicate that our proposed explanation approach is deemed by users as significantly more useful for understanding Bayesian Network Reasoning than another existing explanation method it is compared to.

Keywords

Bayesian Networks explanation, explainable Artificial Intelligence, Natural language explanations, human evaluation of explanations, evaluation in the medical domain

1. Introduction

It is generally accepted that a proper explanation of AI models is one of the requirements for trustworthiness [1, 2, 3]. Whereas the accuracy of an AI model is important in many fields, the inability to explain the reasoning or rationale behind the model may block any perspective on its real-life usage, especially in critical domains. Within AI, Bayesian Networks (BNs) can represent knowledge and perform reasoning in contexts of uncertainty. However, interpreting BNs reasoning may be quite a complex task for users because the reasoning mechanism can run in different directions (e.g., from causes to consequences and vice versa). Moreover, the linkage between variables can lead to complex and indirect relationships, complicating the interpretation.

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain
*Corresponding author.
nikolay.babakov@usc.es (N. Babakov)

BNs are directed acyclic graphs whose edges show causal or influential relationships between nodes and that can encode any discrete or continuous variable (see Figure 1a for an illustrative example of a BN structure). BNs have been used in various fields: medicine [4], law [5], harvesting [6], etc. To complement these efforts, many researchers [7, 8] are now focusing on studying better ways to present the reasoning of BN to domain experts.

In general, BN reasoning is presented either visually [9] or in natural language [8, 7]. In this paper, we focus on textual explanations. Arguably, the most important challenge of textual approaches is content determination [10] - the decision on what information (e.g., what variables in the case of BNs) should be used for an explanation.

To the best of our knowledge, the existing textual explanation approaches do not explicitly consider the path through which the message passes from the evidence node(s) to the target node. Delivering the explanation using such paths could allow BN users to understand the exact chains of reasoning within a BN that yield the final probability update in the target node. In this paper, we consider a novel content determination strategy that finds the most significant of the aforementioned paths and then textually presents them. We refer to these paths as factor arguments. They are delivered to a BN user in natural language. We compare the proposed approach to existing approaches with human-driven evaluation and open-source our code[1].

The rest of the paper is organized as follows. Section 3 outlines all the details of the proposed method. Section 4 describes the design and the result of computational experiments aimed to test the computational limitations of the proposed explanation approach. Section 5 reports the results of the human evaluation of the method in the medical domain.

2. Related work

There are many explanation methods for Bayesian networks [11, 12] that can be delivered in different modalities. Druzdzel and Henrion [13] proposed two types of Bayesian network explanations: Qualitative Belief Propagation and Scenario-based Reasoning. Qualitative Belief Propagation focuses on tracing the qualitative effects of evidence through a belief network, emphasizing the direction and impact of evidence from one variable to the next. Scenario-based Reasoning, on the other hand, generates alternative causal stories to account for the evidence, offering a narrative-based approach to understanding the outcomes of probabilistic reasoning. Both approaches aim to enhance the comprehensibility of Bayesian inference, catering to different aspects of human reasoning under uncertainty. Zukerman et al. [14, 15] explain BNs using an iteratively generated argument graph, which consists of a subgraph of said BN. The Elvira tool [16] highlights the links within the BN that offer qualitative insight into the conditional probability tables. [17] generates quantified statements and reasons with text using fuzzy syllogism. [18] deliver an explanation in a table view, exploiting the generalized Bayes factor score to determine important nodes. [19] generates contrastive explanations using the annotated lattice obtained by the relations of BN nodes, and [20] uses the concept of Maximum A Posteriori independence to define the nodes relevant to the reasoning explanation.

Our approach delivers the explanation as a sequence of text statements describing the process of BN reasoning. It builds on previous work on extracting chains of reasoning from BNs like

[1] https://gitlab.nl4xai.eu/nikolay.babakov/bn_explanation_with_factor_arguments/

the one used in the INSITE system [21]. It was later refined in [22] and [7]. These approaches suggest how to measure and explain the effect of the available evidence on a target node, but they are quite limited when it comes to explaining interactions between chains of reasoning. We also expand the proposal of [23] which explains BNs using argumentation theory. In this work, each conditional probability table is translated into a rule, producing an argumentation scheme that operates similar to Maximum-a-posterior probability inference. We introduce our own formal notion of an argument, that attempts to isolate the effects of different inference paths in the network.

In [24] an initial argumentation diagram is refined and labeled to capture the interactions in a BN. This work led us to propose an alternative characterization of argument independence. In [25] the BN explanatory approach restructures the BN such that the target node has its Markov nodes as parents, then the target CPT is condensed into decision trees, and finally the explanatory nodes under specific contexts are identified by traversing the DTs, and the corresponding BN explanations are generated dynamically. A similar approach was proposed in [9, 26] where an argument diagram named support graph (a directed subgraph of the original BN) is extracted from the original BN and further used for explanation. It relates all available evidence to the target in a series of inference steps. The named approaches do not disentangle the separate effects of the different premises of each step on the rule conclusion, as far as the final explanation is delivered as one argument without explicit designation of what particular evidence node or path of belief update made a more significant impact on the final update of the probability. In our work, we address this limitation.

Unlike modern model-agnostic explainability frameworks like LIME [27] and SHAP [28], we focus on explaining which relations between variables are more important rather than explaining which variables are more important overall.

To generate our explanations, we follow some of the guidelines established in [29]. In particular, we emphasize the need for a selective explanation, offer a contrastive explanation, and show how to explain quantitative arguments in qualitative terms.

3. A proposal for explaining Bayesian Networks

Users normally interact with the BN by entering evidence and querying how the probability of a target variable changes in response. For example, for the BN in Figure 1a the user may indicate that the patient does not have Tuberculosis and has abnormal XRay results and may be interested in learning how likely it is that the patient has Lung cancer.

Our goal in this paper is to propose a model for building natural-language explanations about the reasoning process occurring in a BN. The content to be included in the explanation is determined as a set of directed subgraphs of the BN's factor graph, which we refer to as factor arguments (FAs). Figure 1c shows an example of a *FA* relating the evidence nodes 'XRay Results' and 'Tuberculosis' to the target node 'Lung cancer', Figure 1d shows a textual description of this *FA*.

In this article, we deal exclusively with BNs with discrete categorical value nodes. Continuous ones fall out of the scope of our work, and while this method can in principle be applied to ordinal values, we do not think our approach is particularly well suited to them.

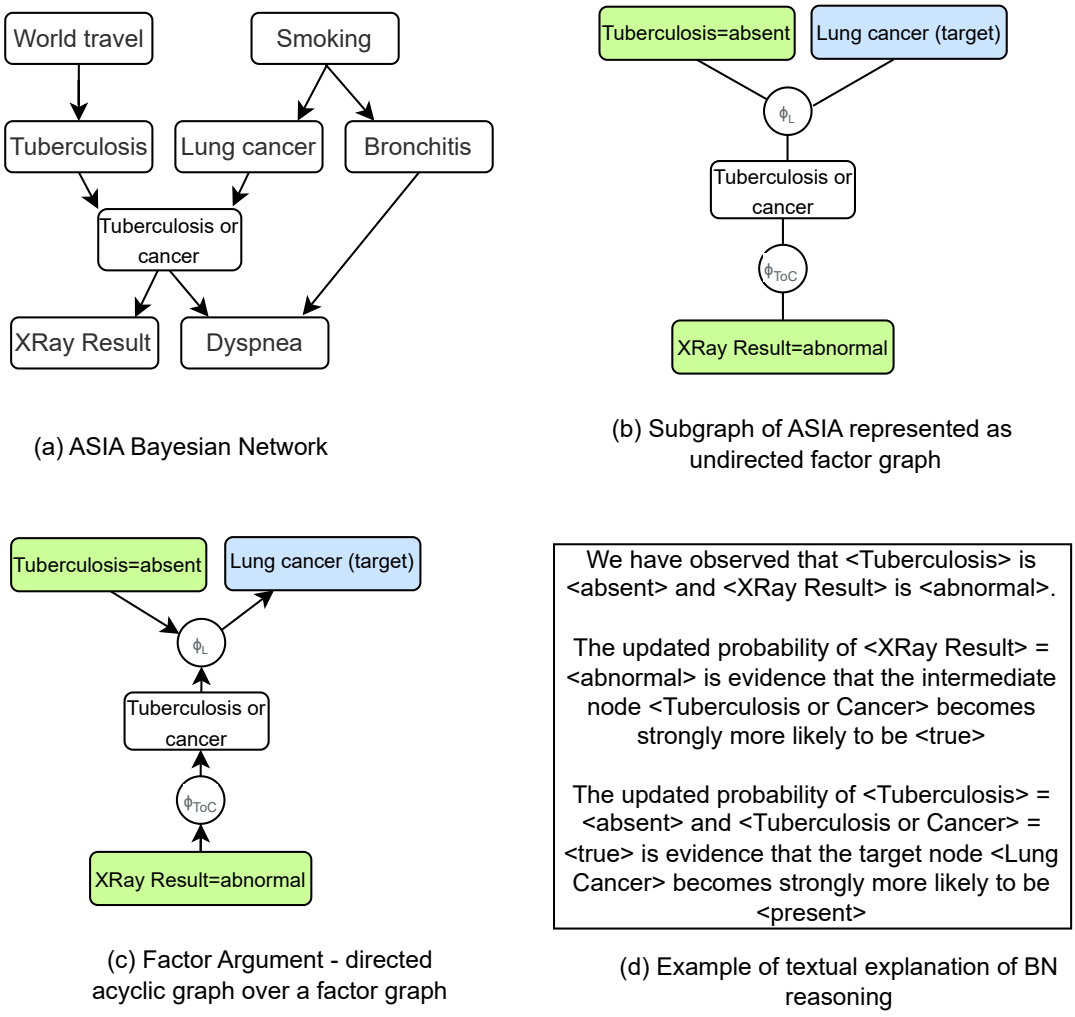


Figure 1: (a) An example of a Bayesian Network is the ASIA Network [30]. (b) Visualization of an undirected subgraph of the ASIA BN factor graph, determining the nodes to be used for the reasoning explanation given two evidence nodes (XRay Result = abnormal and Tuberculosis = absent) and a target node Lung cancer. (c) Visualization of Factor Argument: directed acyclic subgraph of a factor graph BN. (d) Possible textual explanation of reasoning.

3.1. Preliminaries

A BN is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph [31]. Its nodes are associated with conditional probability tables (CPTs), which express the probability of a variable taking a particular value given the outcomes of its direct predecessors.

CPTs may be represented as a factor [31], which is a function mapping all possible values of one or more random variables (its scope) to positive real numbers corresponding to the values

	Tuberculosis	
World Travel	Yes	No
Yes	0.05	0.95
No	0.01	0.99

World Travel	Tuberculosis	ϕ_T
Yes	Yes	0.05
Yes	No	0.95
No	Yes	0.01
No	No	0.99

Table 1
An example of CPT of Tuberculosis node from the ASIA Bayesian Network (left) represented as a factor (right).

in the CPT (see Table 1 for an example). The Scope of a node refers to the set of variables that are directly influenced by or directly influence a given node within a network. Factors can be multiplied, divided, marginalized, and normalized. They can be used to represent a BN as a factor graph - an undirected bipartite graph containing variable and factor nodes. The factor graph only contains edges between variable nodes and factor nodes, and it is parameterized by a set of factors (see Figure 1b). Each factor node is associated with precisely one factor, whose scope is the set of variables adjacent to the factor node in the graph. [2]

Our explanation framework is based on the loopy message passing algorithm for inference in a Bayesian Network [31]. The goal of this algorithm is to approximately compute queries of the form

$$P(T = t|E_1 = e_1, ..., E_m = e_m)$$

That is, given some evidence $E_i = e_i, i = 1, ..., m$ we want to compute how beliefs about a target variable T in the network change. Message passing starts with defining the initial potential factors of each node in the factor graph. The factors of factor nodes are lifted from the respective CPT, and the factors of variable nodes are defined as lopsided factors (where one state probability equals 1 and others 0) if their value is known and as constant factors otherwise.

The information between the variable nodes is communicated with messages of the form $\mu_{A,B}, \mu_{B,A}$ for each pair of adjacent nodes in the associated factor graph. Each message is initialized as a constant factor and aimed to connect variables of the factor graph in both directions. The factors are updated with the following formula:

$$\mu_{A,B} = \sum_{\text{Scope}(A)\setminus\text{Scope}(B)} \phi_A \prod_{C\neq B} \mu_{C,A}$$

Namely, we take the factor ϕ_A associated with the sender node A, multiply it by all messages sent to A, except the one coming from the receiver B, and finally marginalize all nodes not in the intersection of scopes of the sender and receiver nodes. [3]

[2] For the sake of simplicity, we will use capital letter variables such as X, Y, Z to refer to variable nodes in a factor graph, and we will use notation such as ϕ_X to refer to the unique factor node that represents the CPT for node X. We will abuse the notation and also refer to ϕ_X as the factor that represents such a CPT.

[3] Note that loopy message passing is a method of approximate inference, which is not guaranteed to converge.

3.2. Factor arguments

While inference algorithms such as message passing will provide an approximately correct conclusion, it is often hard to understand, even for experts, how the evidence influenced the results. To make the BN reasoning process more understandable, we want to deliver to the user a list of considerations explaining how the given pieces of evidence affected the target variable. To do so, we need a way to abstractly represent the considerations that will be turned into explanations. For this purpose, we introduce factor arguments (*FAs*) - webs of flowing evidence relating the evidence nodes and the target node. We represent *FA*s as directed acyclic graphs over a factor graph, which trace the path followed by messages in loopy message propagation from the evidence to a target node.

Definition 1 (Factor Argument (*FA*))**.** *A factor argument in a BN is a directed acyclic graph whose skeleton corresponds to a subgraph of the factor graph of said BN. As such, it is composed of alternating variable and factor nodes.*

A factor argument has a single sink, corresponding to a variable node, we will call the target node. Each of its sources is also a variable node, which we will call the evidence node.

An example of a factor argument is shown in Figure 1c.

3.3. Step effect and factor argument effect

To define the effect that a *FA* has on the target variable, we introduce the notion of *factor argument effects*. We can understand a *FA* as a series of inference steps, where for each factor node in the *FA* we combine some belief updates about its direct predecessors with the factor itself to produce a belief update for the successor of the factor node. The direct predecessors in *FA* are defined as follows:

Definition 2 (Direct Predecessors in Factor Argument ($Pred_{FA}$))**.** *For any node (either factor node or variable node) within a Factor Argument (FA), the set of direct predecessors, denoted as $Pred_{FA}(node)$, includes all nodes that have a direct edge leading to the considered node within the FA. Formally, for a variable node X within an FA:*

$$Pred_{FA}(X) = \{\phi \mid \phi \text{ has a direct edge to } X \text{ within the FA}\} \tag{1}$$

and for a factor node ϕ :

$$Pred_{FA}(\phi) = \{X \mid X \text{ has a direct edge to } \phi \text{ within the FA}\} \tag{2}$$

Note that the direct predecessors of a factor node will all be variable nodes, and viceversa. Our goal will be to summarize the flow of evidence between variable nodes as mediated by factor nodes. In other words, we will break down a *FA* in a series of steps, one for each factor in the factor argument, where a belief update on its direct predecessors, i.e., its premises, will produce a belief update on its only successor.

In our formalism, similarly to message passing, beliefs about variables are represented as factors. We use the notation δ_X to represent a belief update on node X, i.e., a factor whose scope is the same as all possible values of X. Suppose we are interested in evaluating the effect

of factor ϕ, whose direct predecessors are $\text{Pred}_{FA}(\phi)$, on its successor node X. We measure this through the step effect of ϕ on X.

Definition 3 (Step Effect (*SE*))**.** *Let X be a variable node in a factor argument FA, ϕ a factor node that precedes X and $\Delta_\phi = \{\delta_Y : Y \in Pred_{FA}(\phi)\}$ a set of belief updates for each of the variable nodes that precede ϕ on FA. The step effect (SE) of factor ϕ on X given premises Δ_ϕ is defined as:*

$$SE(\phi, \Delta_\phi, X) = \frac{\sum_{Pred_{FA}(\phi)} \phi \cdot \prod_{\delta \in \Delta_\phi} \delta}{\sum_{Pred_{FA}(\phi)} \phi} \tag{3}$$

An example of calculating a step effect is shown in Figure 2.

The definition of *SE* is reminiscent of the message-passing equations. However, we additionally perform division by the marginalized factor to separate the information derived from the updates Δ_ϕ and the information passively contained in the factor ϕ. Note that the definition of *SE* is purely heuristic, and it does not correspond to any clear semantics. In section 4 we will later show that, in practice, this content determination algorithm approximates loopy message propagation.

Now, having defined the *SE*, we may define the way to compute the factor argument effect (*FAE*) updated beliefs of a complete *FA* on each of the nodes it involves.

Definition 4 (Factor Argument Effect (*FAE*))**.** *Let FA be a factor argument within a Bayesian Network, consisting of a sequence of factor nodes and variable nodes leading from evidence nodes to a target node. The Factor Argument Effect (FAE) of FA on a target node $X \in FA$, denoted as $FAE(FA, X)$ or δ_X, is defined as the cumulative effect of recursively calculating the SE along the FA from the evidence nodes to its target node X. Formally,*

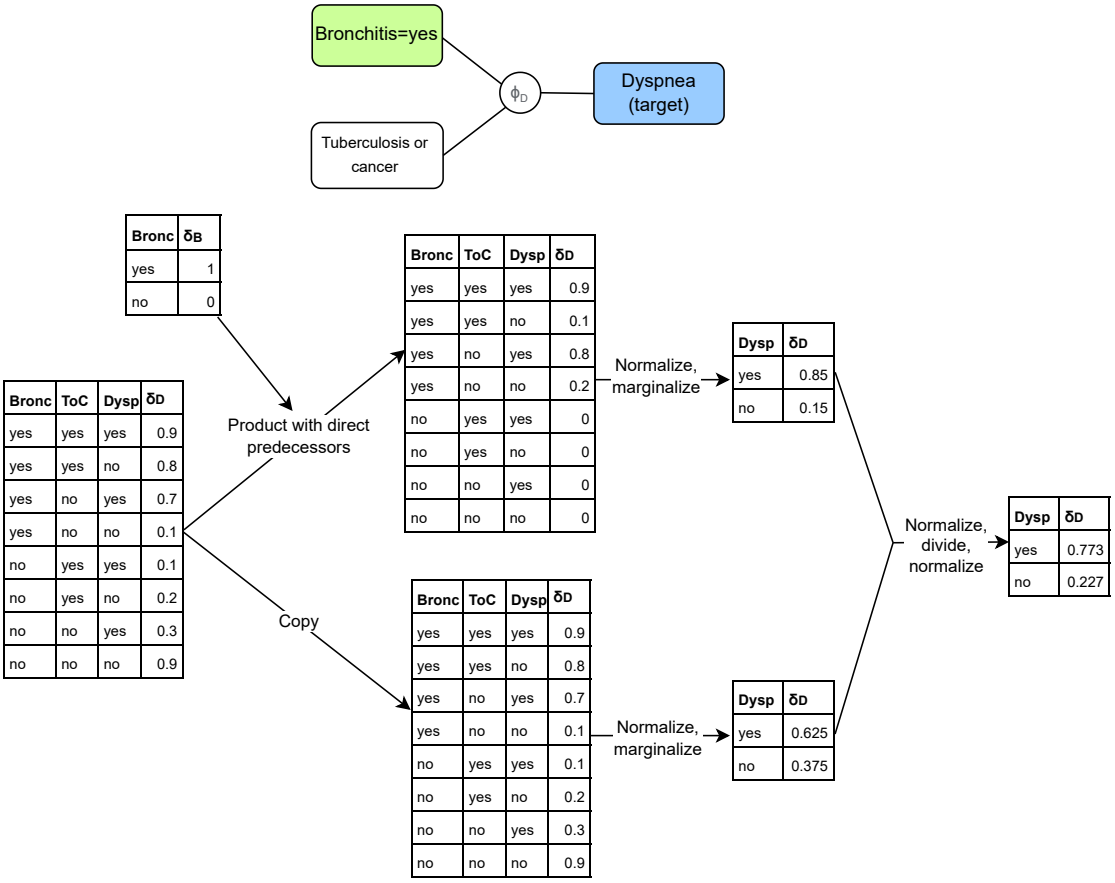
$$\delta_X = FAE(FA, X) = \prod_{\phi \in Pred_{FA}(X)} SE(\phi, \Delta_\phi, X), \tag{4}$$

where $\Delta_\phi = \{\delta_Y : Y \in Pred_{FA}(\phi)\}$ is a set of belief updates δ_Y for each node Y preceding ϕ in the FA. These belief updates are either computed recursively or are lopsided factors representing a piece of observed evidence. More formally,

$$\delta_Y = \begin{cases} \textit{FAE(FA, Y), if } Y \notin \textit{ observed evidence} \\ \textit{Obs(Y = y), otherwise} \end{cases},$$

where Obs is a function that returns a lopsided factor, and Obs(Y = y) is a lopsided factor such as [$True$: 0, $False$: 1] for the observation $Y = False$.

In Figure 3, we show how to compute the *FAE* by recursively applying *SE*, consequently updating the beliefs of all variable nodes within the *FA*. The definition of the *FAE* allows quantifying the effect of *FA* on our beliefs about each of its intermediate nodes. We will be able to build on this capacity later to produce textual explanations of *FAs*.



Bronc	δ_B
yes	1
no	0

Bronc	ToC	Dysp	δ_D
yes	yes	yes	0.9
yes	yes	no	0.8
yes	no	yes	0.7
yes	no	no	0.1
no	yes	yes	0.1
no	yes	no	0.2
no	no	yes	0.3
no	no	no	0.9

Bronc	ToC	Dysp	δ_D
yes	yes	yes	0.9
yes	yes	no	0.1
yes	no	yes	0.8
yes	no	no	0.2
no	yes	yes	0
no	yes	no	0
no	no	yes	0
no	no	no	0

Dysp	δ_D
yes	0.85
no	0.15

Bronc	ToC	Dysp	δ_D
yes	yes	yes	0.9
yes	yes	no	0.8
yes	no	yes	0.7
yes	no	no	0.1
no	yes	yes	0.1
no	yes	no	0.2
no	no	yes	0.3
no	no	no	0.9

Dysp	δ_D
yes	0.625
no	0.375

Dysp	δ_D
yes	0.773
no	0.227

Figure 2: Visualization of Step Effect calculation. The numerator is calculated by multiplying the initial factor by incoming evidence factors, then normalizing, and marginalizing by the direct predecessor nodes. The denominator is calculated by marginalizing by the direct predecessor nodes and normalizing. After division, the resulting factor is normalized. Note that for compactness of visualization, we replaced explicit names of node states with "yes" and "no" notation. "ToC", "Bronc", and "Dysp" are the abbreviations of the corresponding nodes.

3.4. Factor argument strength

The definition of the *FAE* is a comprehensive description of how the *FA* affects our beliefs of all nodes from evidence to target variable nodes. But if we want to only show to the user those *FAs* that are most relevant, we need to define the way of comparing *FA* effects to each other. For this, we define the notion of *factor argument strength* (*FAS*) w.r.t. a value of the target variable $T = t_o$ as follows:

Definition 5 (Factor Argument Strength (*FAS*))**.** *Given a Bayesian Network and a factor argument FA affecting the belief state of a target variable T, the Factor Argument Strength (FAS) with respect to a particular value of T (denoted as $T = t_o$) is defined as follows:*

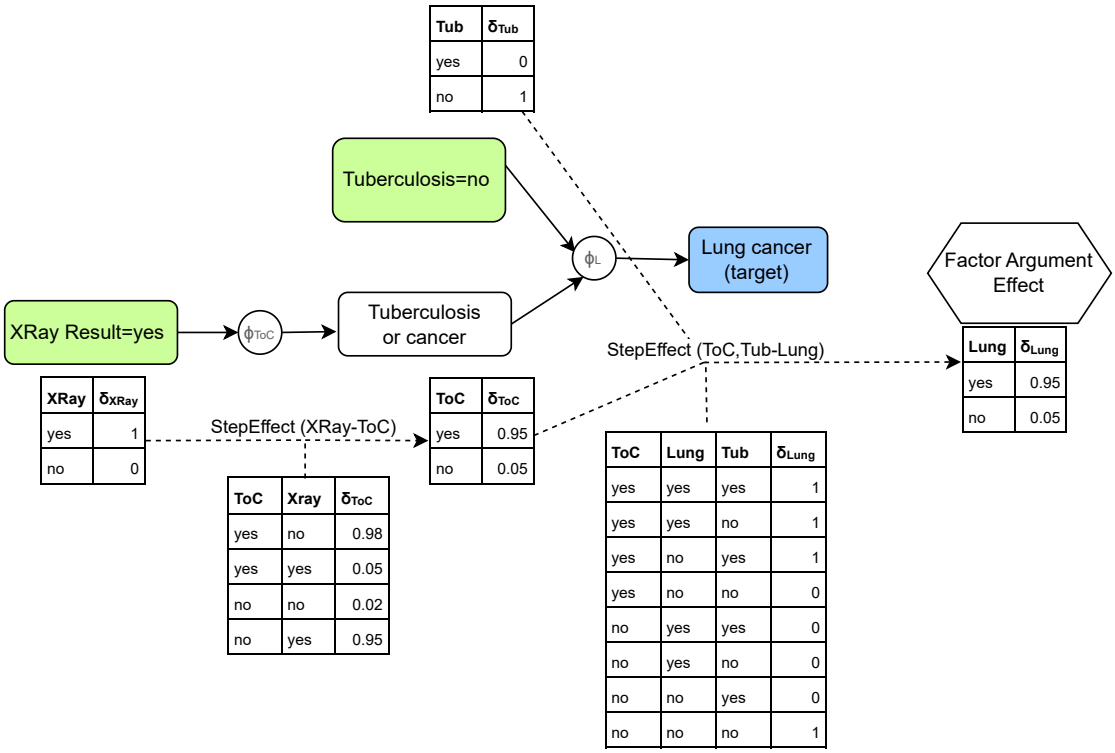


Figure 3: Visualization of Factor Argument Effect calculation. Step Effect is recursively calculated from evidence nodes to target node consequently updating the beliefs of all variable nodes within the Factor Argument. Note that for compactness of visualization, we replaced explicit names of node states with "yes" and "no" notation."ToC", "Tub", and "Lung" are the abbreviations of the corresponding nodes.

$$FAS(FA, T = t_o) = \log \frac{\delta_T(t_o)}{\frac{1}{N-1} \sum\limits_{i \neq o} \delta_T(t_i)} \tag{5}$$

where:

- N *is the number of possible states of the target variable* T*,*
- $\delta_T(t_o)$ *is the Factor Argument Effect associated with the target variable taking the value* t_o *due to the factor argument* FA*,*
- $\delta_T(t_i)$ *is the Factor Argument Effect associated with all other states* t_i *of* T*, excluding* t_o *due to the factor argument* FA*.*

Our definition of *FAS* satisfies one intuitive property: if the target node in the corresponding subgraph of the Bayesian Network associated with *FA* is d-separated conditional on the evidence nodes of the FA, then its associated *FAS* is zero. This can be seen from Eq. 3 – if the reasoning step from the direct predecessors $\text{Pred}_{FA}(\phi)$ is d-blocked, then the equation simplifies to a constant factor, which propagates along its path. In Eq. 4, since all paths are d-blocked by

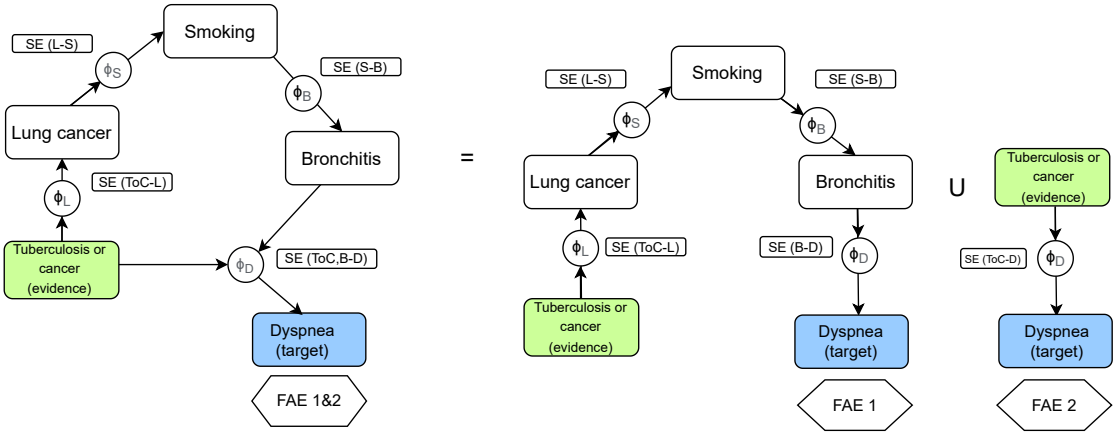


Figure 4: The complex factor argument (FA) can be decomposed into two simpler *FAs*. This decomposition is a good approximation if the factor argument effect (FAE) of the complex argument is approximately similar to the product of the *FAEs* of the simpler arguments. Arguments of *SE* (ToC, L, S, etc) are the abbreviations of the corresponding nodes.

assumption, the final node of all incoming *SE* is constant. Finally, in Eq. 5 the strength of a constant factor is 0.

The idea behind *FAS* is inspired by [32], where the likelihood ratio of the hypothesis given the evidence was used to quantify the strength of a chain of reasoning. We, however, use a notion of strength directly based on the CPTs of the BN, which allows us to talk about the strength of specific statistical relations rather than the strength of (a subset of) the evidence variables. We can use the *FAS* to rank different *FAs*, and select those that are most important to defend a given conclusion.

3.5. Splitting factor arguments

During interaction with the BN, multiple pieces of evidence may be provided. Moreover, BNs may involve multiple simple paths between even a single evidence node and a target node. Consider an example in Figure 4, where a single evidence node "Tuberculosis or Cancer" may be connected through two different simple paths to the target node "Dyspnea".

This fact raises the question whether these two *FAs* should be delivered to the BN user jointly or separately. This problem was previously discussed, e.g., in [24], where the author explains the distinction between convergent arguments — where each of them independently supports a conclusion — and linked arguments — where the strength of each *FA* depends on the presence of the other.

Thus, we need a formal way to decide whether the *FAs* are independent. If they are, then we can break down the composite *FA* into its simpler subcomponents. But if the interaction between the *FAs* is important to support the conclusion, we need to present the more complex *FA* to the user. The formal definition of independent FAs is as follows:

Definition 6 (Independent Factor Arguments)**.** *Two factor arguments (FAs) within a Bayesian*

Network are considered independent if, and only if, the product of the Factor Argument Effects (FAEs) of these separate FAs equals the FAE of their combined effect, where the combination of FAs refers to the union of their corresponding graphs. Formally, let FA_1 and FA_2 be two distinct FAs, and let FA_{union} denote their union. The FAs FA_1 and FA_2 are independent if:

$$FAE(FA_1) \times FAE(FA_2) = FAE(FA_{union}) \tag{6}$$

As an illustration, we can consider what would happen in an AND network. This network has a $A \rightarrow C \leftarrow B$ structure where A and B nodes are distributed as a Bernoulli$(1/2)$, and C is true *iff* A and B are true. The *FAE* of the argument $A = 1$ on C is $(C = 1 : 3, C = 0 : 1)$, and similarly for $B = 1$. However, the factor effect of the joint *FA* $A = 1, B = 1$ on C is $(C = 1 : 1, C = 0 : 0)$. This is different from the product of the individual effects, so we conclude that any *FA* we present to the user must present both *FAs* jointly. See the implementation of this example in the demonstration notebook in our repository (see Section 1).

In terms of explanations, the definition of independence is too restrictive - *FAs* with weak interaction with one another should still be presented independently to the user. Thus, we define the notion of approximate independence of *FAs* as follows:

Definition 7 (Approximate Independence of Factor Arguments)**.** *A set of Factor Arguments (FAs) within a Bayesian Network is said to be approximately independent if the product of the Factor Argument Effects (FAEs) of any subset of these FAs is within a specified threshold of distance from the FAE of the union of FAs in said subset. Formally, for a given threshold θ, a set of FAs $\{FA_1, FA_2, \ldots, FA_n\}$ is approximately independent if for any non-empty subset $\{FA_i\} \subseteq \{FA_1, FA_2, \ldots, FA_n\}$, the following condition holds:*

$$\left| FAE(\bigcup_{FA_i} FA_i) - \prod_{FA_i} FAE(FA_i) \right| < \theta, \tag{7}$$

where $\bigcup_{FA_i} FA_i$ denotes the union of the factor arguments in the subset, and $FAE(FA_i)$ represents the Factor Argument Effect of FA_i.

This notion captures the degree to which the combined effect of a set of FAs deviates from what would be expected if they were truly independent, based on their individual effects.

To decide whether the effect of the union of arguments is close to the product of the effects of the individual arguments we introduce the notion of factor argument distance (*FAD*). The formal definition of measuring the distance between two *FAE*s combines the ideas of *FAE* and *FAS*. The *FAE* inferred with a certain *FA* about the target variable δ_T is expressed as a factor. To quantify the difference between two δ_T factors, we may divide one factor by another and then look for the biggest resulting *FAE* from all possible states of the target within the divided factor. The formal definition of *FAD* is as follows:

Definition 8 (Factor Argument Distance (*FAD*))**.** *The Factor Argument Distance (FAD) between two factor arguments (FAs) within a Bayesian Network, considering their effects on the belief state of a target variable T, is defined through the differential impact of these FAs, quantified by their FAEs. This is captured by the equation:*

$$FAD(\delta_T^a, \delta_T^b) = \max_{i<N} |\log \frac{\delta_T^{a/b}(t_i)}{\frac{1}{N-1}\sum_{j\neq i}\delta_T^{a/b}(t_j)}| \tag{8}$$

where δ_T^a, δ_T^b are two different belief updates, $\delta_T^{a/b}(t_i)$ is a shorthand for $\delta_T^a(t_i)/\delta_T^b(t_i)$ and we assume that variable T has N possible outcomes numbered from 0 to $N-1$.

Intuitively, the definition of *FAD* corresponds to identifying the outcome t_i on which factors δ_T^a, δ_T^b disagree the most, and then quantifying in terms of logodds the difference between the respective updates they would induce if applied to variable T.

This definition yields small values if the *FAE* about the target variable δ_T inferred by two different *FAs* are similar, and bigger values otherwise. This is fairly easy to see, since if $\delta_T^a \approx \delta_T^b$ then we will have that $\delta_T^{a/b}(t_i) \approx 1$, and so $FAD(\delta_T^a, \delta_T^b) = \max_{i<N} |\log \frac{\delta_T^{a/b}(t_i)}{\frac{1}{N-1}\sum_{j\neq i}\delta_T^{a/b}(t_j)}| \approx \max_{i<N} |\log \frac{1}{\frac{1}{N-1}\sum_{j\neq i} 1}| = 0$. For dissimilar factors, the FAD will be above zero. As desired, we can use this to decide if the effect of the union of arguments is close to the product of the effects of the individual arguments.

We finally define a *FA* to be approximately proper:

Definition 9 (Approximately Proper Factor Argument)**.** *A Factor Argument (FA) within a Bayesian Network is deemed approximately proper if it cannot be decomposed into a union of approximately independent FAs. This condition implies that the FA, as a whole, contributes uniquely to the inference process without being reducible to simpler, approximately independent components. Formally, an FA is approximately proper if, for any partition of the FA into subsets $\{FA_1, FA_2, \ldots, FA_n\}$, where each FA_i is an approximately independent factor argument, there exists no combination of these FA_i's whose union effectively replicates the inference impact of the original FA within a specified approximation threshold.*

This definition ensures that an *FA* is considered approximately proper only if its contribution to the Bayesian inference cannot be approximated by combining the effects of smaller, simpler arguments, thereby maintaining the integrity and the unique inferential value of the *FA* within the network's reasoning process.

Our next goal will be to identify all the proper *FAs* relating the available evidence to a target. We explain how to do this in the next section.

3.6. Core algorithm overview

At this stage, we are ready to define our content determination approach for the BN reasoning explanation. Its idea is to construct a set of proper, maximal, and independent *FAs* relating the evidence to the target node. Note that maximal *FAs* are defined as *FAs* that are not a subgraph of another *FA*. The pseudocode of the algorithm is shown in Algorithm 1.

The inputs to the algorithm are the factorized BN, the target node, and the observed evidence. First, we construct all *FAs* corresponding to simple paths from evidence nodes to the target

Algorithm 1 Overview of the content determination algorithm for Bayesian Network explanation

function FindMaximalProperFAs(BN, t, E)
  Input: BN ▷ Bayesian Network in factor view
  Input: t ▷ target node
  Input: E ▷ evidence nodes
  Output: $ProperAPs$ ▷ list of independent factor arguments
  Constant ML ▷ max length of the simple path
  Constant MC ▷ max number of simple paths
  Constant DT ▷ Dependence Threshold
  AllSimplePaths = GetAllSimplePaths(BN, t, E, ML)
  for $SPCombination \in Combinations(AllSimplePaths, MC)$ **do**
    IsDependent = CheckDependence(SPCombination,DT)
    if IsDependent **then**
      Add ComposeFAs(SPCombination) to ProperFAs
  ProperFAs = AdjustProperFAs(ProperFAs)
  OutputFAs = FilterNonMaximalFAs(ProperFAs)
  if ML $< \infty$ or MC $< \infty$ **then**
    OutputFAs = PairwiseCombine(OutputFAs)
      ▷ Guarantees pairwise independence when heuristics are applied
  return OutputFAs
function CheckDependence($SPCombination, DT$)
  SPCombinationUnion = ComposeFAs(SPCombination)
  SPCombinationUnionEffect = FAEffect(SPCombinationUnion) ▷ Eq. 4
  for $SPCombinationUnionPart \in Partitions(SPCombination)$ **do**
    SPCombinationUnionPartEffect =
      FAEffect(SPCombinationUnionPart)
    Distance = FADistance(SPCombinationUnionEffect,
      SPCombinationUnionPartEffect) ▷ Eq. 8
    if Distance $<$ DT **then return** False
  return True
function AdjustProperAPs($ProperAPs$)
  ProperAPs = RemoveSubgraphs(ProperAPs)
  ProperAPs = PairwiseRefinement(ProperAPs)
    ▷ Combine non-independent pairs of FA until all FA are independent
  return ProperAPs

node. Each possible combination of simple paths passes the dependence check, which verifies whether the corresponding composed *FA* is proper in the *CheckDependence* function (note that every FA can be expressed as a combination of simple paths, so this loop exhausts all FAs). To calculate this, we compare the *FAE* of the composed *FA* to the product of the *FAE* of the union of each possible partition of the arguments that compose it using *FAD* between these *FAE*. If the *FAD* between the *FAE* of any possible partition and the *FAE* of the composed *FA* is below the given threshold (parameter *DT*), then we conclude that the current combination of *FAs* is independent; hence, it is not to be delivered jointly. Otherwise, if the *FAD* of all possible partitions is always above the threshold, we conclude that the composed *FA* is proper and will

be delivered jointly. Finally, we filter for maximal *FAs*. The result is a set of maximal, proper, and independent *FAs* that capture how the evidence relates to the target.

Note that we have omitted two optimizations in the *CheckDependence* function for simplicity of exposition. First, we only check for partitions composed of *FAs* previously identified to be proper. Second, the *FAEs* of each composed argument are stored and dynamically reused between calls. To facilitate these optimizations, we need to iterate from *SPCombinations* of fewer *FAs* to those of more *FAs*. The implementation details are available in our repository (see Section 1).

The result of the proposed algorithm is the list of *FAs* that will be delivered to the user, separately and in descending order w.r.t. their *FAS* (Eq. 5). Typically, we show only a handful of these *FAs* - namely those that have the highest strength and thus are most relevant for the user. We can control how many *FAs* to show by filtering those with *FAS* below a threshold and/or only showing the top N *FAs*. Once we have selected the *FAs* to show to the user, we need to explain them in natural language.

3.7. Ways to overcome algorithm limitations

The proposed algorithm has certain limitations. Since the number of simple paths in a graph is factorial to the number of variables and the number of complex *FAs* is exponential on the number of simple paths, the naive approach of listing all *FAs* and computing all effects will be impractical for big networks. Some heuristics can relieve these issues. First, we can consider only simple paths with lengths below a threshold (parameter *ML*). Second, we can only consider complex *FAs* combined from a limited number of simple *FAs* (parameter *MC*).

These heuristics void the guarantee that the output will be a set of independent *FAs* since some *FAs* combinations will not be tried. To alleviate this issue, when applying the heuristics, we iteratively combine pairs of dependent *FAs* to guarantee pairwise independence. This procedure removes the guarantee that the constructed *FAs* will be proper. The last steps to verify that the selected *FAs* are proper are to verify that none of these *FAs* is a subgraph of another and to make sure that none of the *FAs* are mutually dependent.

3.8. Textual explanation with extracted factor arguments

Having identified the set of *FAs* to be presented, the final step is explaining these *FAs* in natural language. Refer to Table 2 for the examples of the different types of the explanation. The details of these explanation types are described below.

3.8.1. Explanation of Factor Argument steps

Overall, the explanation of *FA* is aimed at verbally guiding the BN user from the evidence node or nodes to the target node using the combination of the templates. The explanation of each step within *FA* could include the explanation of observations, inference rules, and conclusion.

Each *FA* starts with a description of the evidence nodes. The description of the evidence is performed using the following template: *We have observed that {evidence node} is {evidence node state}* (e.g., “We have observed that <XRay Result> is <abnormal>”).

Type	Text of explanation
Overview	Since <XRay Result> is <abnormal> and <Tuberculosis> is <absent>, we infer that <Lung Cancer> = <present>.
Direct	We have observed that <XRay Result> is <abnormal> and <Tuberculosis> is <absent>. The updated probability of <XRay Result> = <abnormal> is evidence that the intermediate node <Tuberculosis or Cancer> becomes strongly more likely to be <true>. The updated probability of <Tuberculosis or Cancer> = <true> and <Tuberculosis> = <absent> is evidence that the target node <Lung Cancer> becomes strongly more likely to be <present>.
Contrastive	We have observed that <XRay Result> is <abnormal> and <Tuberculosis> is <absent>. The updated probability of <XRay Result> = <abnormal> is evidence that the intermediate node <Tuberculosis or Cancer> becomes strongly more likely to be <true>. Usually, if the <Tuberculosis or Cancer> = <true> then the <Lung Cancer> = <true>. Since the <Tuberculosis> is <absent>, we can be strongly more certain that <Lung Cancer> = <true>.

Table 2
Three types of converting an *FA* from Figure 1 into textual form.

After the evidence nodes are verbalized, all next steps until the target node include the description of premises that describe the updated beliefs of the nodes w.r.t. the previous *FA* step, inference rules applied on the current *FA* step, and the conclusion from the application of the rules.

The premises are described using the following template: *The updated probability of {previous node in FA} = {state of previous node in FA }* (e.g., "The updated probability of <XRay Result> = <abnormal>"). The exact state of the node is selected to be verbalized if it is the state of the evidence node, or if this state has the highest magnitude of the logodds update.

The description of inference rules and their conclusions slightly varies according to the reasoning patterns applied at a certain step within *FA*: evidential, causal, or intercausal [31].

If the reasoning pattern of the particular step within *FA* corresponds to causal or evidential reasoning, we use the following template, which will be further referred to as direct explanation: *{premises} {verb} {description of the conclusion with a strength qualifier}* (e.g., "The updated probability of <XRay Result> = <abnormal> is evidence that the intermediate node <Tuberculosis or Cancer> becomes strongly more likely to be <true>").

The verb is either *"causes"* or *"is evidence that"* for causal and evidential reasoning, respectively. The strength qualifier reflects the magnitude of beliefs logodds change according to the following mapping: *"Certainly"*, *"Strongly"*, *"Moderately"*, *"Weakly"*, and *"Tenuously"* for cases when the range of logodds update is greater than 10, 1, 0.5, 0.1, and less than 0.1, respectively.

If the reasoning pattern of the particular step within *FA* corresponds to intercausal there could be two options of explanation: direct and contrastive. In the direct explanation, the description is the same as in the purely evidential case we described above. In the contrastive explanation, we compute the counterfactual effect we would have gotten had we observed the premise corresponding to the child of the conclusion node, but not the co-parents (the node or

nodes indirectly connected through a common child node with a target node), and we compare that to the actual inference.

The template of the contrastive explanation consists of two parts. The first part is: *Usually, if {child potential premises} then {counterfactual outcome description}* (e.g., "Usually, if the <Tuberculosis or Cancer> = <true> then the <Lung Cancer> = <true>").

It describes the effect we would have gotten had we observed the premise corresponding to the most likely state of the conclusion node's child, where the state of the child node (child potential premises) and the state of the target node within current *FA* step (counterfactual outcome description), which become more likely within the *FA* are described in plain text.

The second part of the template is *Since {co-parent's premises}, {factual outcome description}* (e.g., "Since the <Tuberculosis> is <absent>, we can be strongly more certain that <Lung Cancer> = <true>").

The co-parent's premises are verbalized in plain text (similar to the first part of the template), while the factual outcome description varies according to the results of the counterfactual outcome. If the most likely state of the target node corresponding to the counterfactual outcome contradicts the state inferred from the actual outcome of the *FA* step, then the factual description is verbalized as *we infer {conclusion description} instead*, where the conclusion description is just a plain text description of the most likely state of the target node within current *FA* step. If the counterfactual outcome corresponds to the actual outcome, we use the template *we {verb} be more certain that {actual outcome description}*, where verb *"can"* is used if logodds of the most likely state of target node inferred from actual inference are higher than those inferred from counterfactual inference. Otherwise, if the actual inference logodds are not greater than the counterfactual ones, the verb *"can not"* is used.

Overall both parts of the explanation could be as follows: "Usually, if the <Tuberculosis or Cancer> = <true> then the <Lung Cancer> = <true>. Since the <Tuberculosis> is <absent>, we can be strongly more certain that <Lung Cancer> = <true>").

Our approach implies that the *FA* may contain multiple simple paths; thus, it is necessary to define the way to convert them into text. We opt for chaining together the explanation of each rule inferred by the factor node and then adding an extra line explaining the cumulative effect of all the rules (e.g., "The updated probability of <XRay Result> = <abnormal> is evidence that the intermediate node <Tuberculosis or Cancer> becomes moderately more likely to be <true>. The updated probability of <Dyspnea > = <present> is evidence that the intermediate node <Tuberculosis or Cancer> becomes moderately more likely to be <true>. All in all, the intermediate node <Tuberculosis or Cancer> becomes strongly more likely to be <true>").

3.8.2. Explanation of the whole Factor Argument

We define three modes of explanation of the whole *FA*: direct, contrastive, and overview. The direct mode verbalizes all *FA* evidence and intermediate steps using the direct template defined above. The explanation in contrastive mode is equal to direct except for the intercausal reasoning *FA* steps, where the contrastive template explanation is used. Finally, the overview produces a simple description of the observations and the conclusion of the *FA* without verbalizing any intermediate steps within the *FA*. Refer to the Table 2 for all three types of explanations corresponding to the *FA* from Figure 1.

4. Experimental results

Our definition of *FAE* is, in a loose sense, meant to imitate the loopy message-passing inference algorithm. We can empirically check the quality of the approximation by comparing beliefs about certain target nodes given certain evidence node values calculated by the message-passing algorithm and our approach.

To perform this comparison, we need to define the way of calculating the belief given the prior belief and *FAE*. Let $\mathcal{FA}$ be the set of relevant, independent *FAs* found by our algorithm w.r.t. the evidence nodes E and target nodes t. The posterior is calculated as follows:

$$\hat{O}(t|\mathcal{FA}) = O(t) \cdot \prod_{FA \in \mathcal{FA}} FAE(FA, t)$$

where $O(t)$ is a factor representing the prior probability of the target variable as computed by the message passing algorithm. This equation, namely, defines the way of approximating the posterior probability of the target node as the product of the prior probability and the *FAE*s of each relevant *FA* on the target.

To study the quality of the approximation, we take BNs of different sizes (from 5 to 37 nodes), collected from the bnlearn website [33], and perform 200 iterations of comparison. Each iteration includes a random choice (with a random seed corresponding to the iteration number) of the evidence nodes and target nodes and further comparison of the probabilities inferred by our algorithm and message-passing algorithm. The main intuition of the definite BN selection was to start the experiments with the smallest BNs and gradually increase the size and treewidth of the BN until the computation time per iteration starts being unreasonable. This selection approach yielded seven BNs with a maximum size of 37 nodes and treewidth equal to 4.

We show the results of our experiments in Figure 5 [4], where we report the mean probability error and correlation between message passing and our algorithm, as well as the average computational time. We report these values w.r.t. the BN nodes number (because this may give an initial intuition about the size of the BN) and also w.r.t. the treewidth because it is known that the complexity of BN inference grows with the treewidth of the BN's graph [34]. To make the calculation faster, we limited the maximal complexity (MC parameter) of the *FA* up to 2. The mean error between probabilities inferred by message passing and our algorithm is somewhat large, varying from 0.07 to 0.14. However, the Spearman correlation between these probabilities is rather high: 0.92 and above (the p-value for all of these cases was significantly less than 0.05). Both, mean error and correlation, do not have a visible trend w.r.t. the increase of either BN nodes number or treewidth. This suggests that the *FAE* approximation of the message-passing process is qualitatively correct.

Figure 6 further illustrates the correlation between the two methods. The logodds implied by our algorithm and the logodds computed through message passing are tightly correlated. However, the slope of the correlation is less than 1, suggesting that our algorithm overweight the strength of the argument. Further work could look into understanding why this is, and how to adjust the strength of the arguments to correct it. Nevertheless, we think that these results suggest that the algorithm is qualitatively focusing on the right parts of the explanation.

[4]The calculations for this figure are performed with an Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz

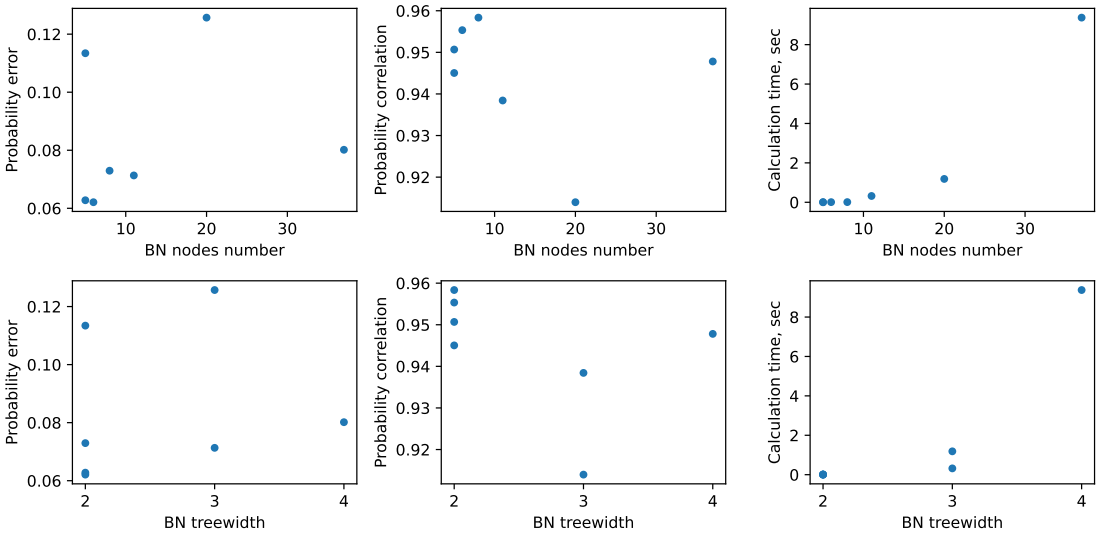


Figure 5: Mean absolute probability error of our approximation vs message passing, Spearman rho correlation coefficient, and calculation time from 200 runs of our algorithm with random target and evidence nodes with different BNs from bnlearn website (cancer, earthquake, survey, asia, sachs, child, alarm). We express the relations both in terms of the total BN node number and the treewidth. The calculations are performed with MC = 2.

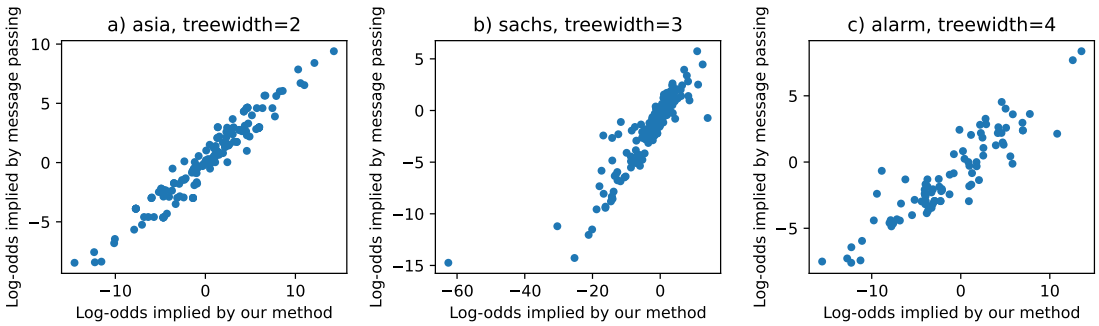


Figure 6: Correlation between the implied log-odds of the query according to our algorithm (x-axis) vs message passing (y-axis). We show the results for three BNs with different treewidth. The calculations are performed with MC = 2.

However, Figure 5 also shows the main drawback of the approach: even limiting the complexity of the *FAs* considered, the processing time increases significantly with the increase in the BN nodes number and its treewidth. Thus, we conclude that the proposed approach may rather precisely imitate the exact inference of the message passing algorithm; however, it is rather to be used for small or medium-sized BNs (around 20 nodes and treewidth equal to 2), and using it with bigger BNs requires further optimization of the algorithm.

We also collect the statistics of certain properties of the proper *FAs* generated with our

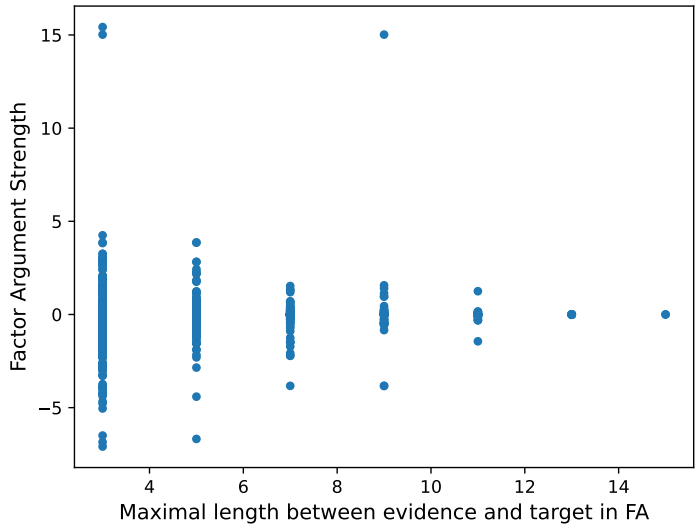


Figure 7: The statistics of proper *FAs* length and *FA* strength from 200 runs of our algorithm with random target and evidence nodes with different BNs from bnlearn website. The calculations are performed with MC=2.

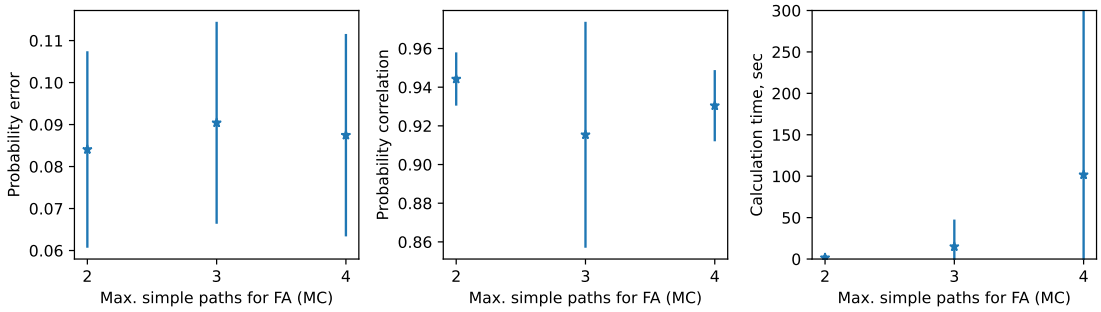


Figure 8: The statistics from 200 runs of our algorithm with random target and evidence nodes with different BNs from bnlearn website related to different values of the MC parameter (the maximum complexity of the resulting *FA*).

algorithm. In particular, we check whether the shorter *FAs* are more often resulted into proper ones. To check this, similarly to the previous experiments, we run *FA* calculations on the BNs from bnlearn website with random evidence and target nodes and plot the statistics of *FA* length and corresponding *FA* strength. The *FA* length is calculated as a maximal path between evidence and target nodes with given *FA*. Figure 7 confirms the intuition that the shorter *FAs* are more often considered proper by our algorithm and also tend to have greater *FAS* than longer *FAs*.

Whereas limiting the complexity of *FAs* may be used as a heuristic for decreasing calculation time, it is necessary to make sure that the *FAs* generated under these heuristics are still qualitatively correct. As far as we have seen from Figure 7 our algorithm is prone to generating shorter *FAs* it seems that the heuristic of limiting maximal complexity of the presented *FAs* (MC

parameter) is more applicable than limiting the maximal length of *FAs* (ML parameter). Thus we study the dynamic of error and correlation of posterior probabilities between our algorithm and message passing algorithm, as well as the computational time with the corresponding standard deviation w.r.t. the maximal complexity of *FA* (MC parameter). Figure 8 shows that there are no significant differences between error and correlation within different values of *FA* complexity. At the same time, the computational time increases significantly with the increase of MC. Thus it seems practical to use MC equal to 2 for practical usage of the proposed algorithm because it yields reasonable probability error and computation time.

5. Human-driven evaluation of the explanation method in the medical domain

In this section, we compare our algorithm with two alternative algorithms using human-driven evaluation relying on widely-known BN describing lung diseases, as well as some of their potential causes and consequences, which was originally composed for demonstration purposes (see Figure 1).

5.1. BN explanation methods for comparison

The first compared method, referred to as a *baseline* simply verbally describes how the probability of the target node is updated given the provided evidence. The textual information is accompanied by graphical tips that include two screenshots of the BN from Netica software[35] before and after the evidence is provided. This idea is similar to the parts of evaluation pipelines in [7, 36] where simply showing BN without clarification was used among other explanation approaches.

The alternative textual explanation approach we perform the comparison to, referred to as *incremental*, was described in [7]. Its main idea is to distinguish the evidence that is considered important between supporting and non-supporting the final change in the target variable. The explanations are accompanied by Netica screenshots highlighting the nodes mentioned in the explanation. We find this method the most suitable for direct comparison because, to the best of our knowledge, this is the only previously proposed algorithm that can be used for textual explanations of any type of BN inference with the ability to set arbitrary nodes both as target and as evidence nodes.

Our method is referred to as *fae*. It also delivers a textual explanation using direct explanation mode together with a visual aid in the form of Netica screenshots with the additional arrays of the *FAs*' direction. Examples of the interface of all explanation methods used for human evaluation can be found in Figures 10,11,12,13 in A.

5.2. Evaluation setup

In general, the human-driven evaluation of BN reasoning explanations is a complex task. To the best of our knowledge, there is currently only one study explicitly dedicated to this task [36]. Moreover, the novel explanation methods, when presented, are not generally compared to the existing ones or demonstrated to the potential users of the explanation [26, 18, 19, 5] with only

rare exceptions, like in [7], where the proposed explanation method was demonstrated to the clinicians.

The core idea of the proposed evaluation setup is to score the explanations first individually and then jointly. When scored individually, the participants are asked whether it is easy to follow the explanation and whether the explanation helped them understand how and why the probabilities in the target node were updated. Five answers from "Strongly disagree" to "Strongly agree" are proposed for selection. When scored jointly, the participants are first asked to choose the method that helped them better understand the reasoning process and then whether the combination of the proposed methods could help. For both joint questions, the participants have the option to answer "None" and "The combination will not help much" respectively. In both individual and joint scoring interfaces, the participant is provided with the option to leave textual feedback. See examples of all questions' interfaces (see Figures 14,15 in A).

At the beginning of the explanation, the participant is presented with BN intuition, including a demonstrative example from Netica and bayesserver [37]. As a reference, we use the ASIA model (see Figure 1). The evaluation includes three scenarios, each corresponding to either predictive, evidential, or intercausal reasoning. To decrease the cognitive load, we randomly present only two scenarios to each participant, making sure that all scenarios get an equal number of demonstrations. Within each scenario, the *baseline* explanation is always demonstrated first, and the *fae* and *incremental* explanations are demonstrated in a mixed order to prevent any order-specific bias.

We engage 25 anonymous participants from our professional network specializing in computer science and engineering research or development. Even though ASIA BN is related to the medical field, it was originally composed for demonstration purposes, so it may be properly understood without specialized knowledge in the medical area. None of them were aware of either which explanation method is ours or any other details of this study. 7 participants had not heard about BNs before the survey; 16 had only a general idea about BNs and only 2 identified themselves as experienced users of BNs. We take certain measures to make sure that the task is understood correctly and that the answers reflect the real properties of the perception of each particular participant. We analyze the textual comments from the participants and the time the person spent on each page.

5.3. Evaluation results

The proposed evaluation design turned out to be rather time-consuming and required a lot of cognitive power from participants. Its median passing time was 25 minutes. We manually examined both quality-control parameters described in the previous section and, finally, as an exclusion criteria, we dropped the answers from 4 participants whose textual comments indicated that they had not understood the task or who spent less than 30 seconds on any page of the task.

The aggregated evaluation results are shown in Figure 9. The "Individual score" plot shows the aggregated answers to the questions about each explanation type. The textual answers are mapped to numerical values from 1 ("Strongly disagree") to 5 ("Strongly agree"). To verify the significance of the difference between the scores, we use a T-test with the null hypothesis of

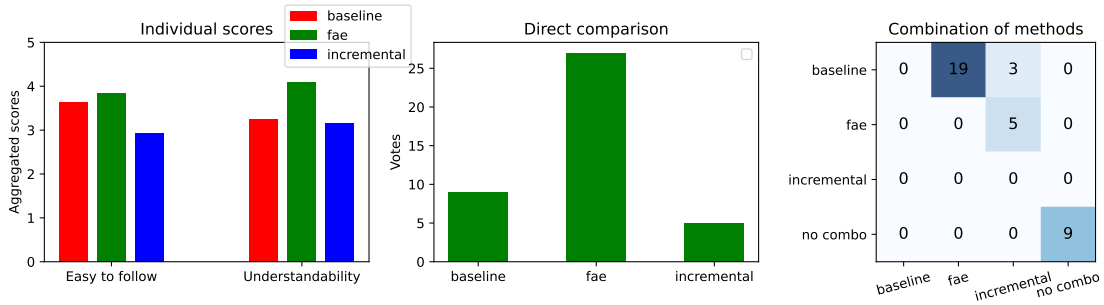


Figure 9: The results of the human-driven evaluation. The "Direct comparison" plot aggregates the answers about the explanation that helped to better understand the reasoning process. The "Combination of methods" plot shows the votes for the question about the potential use of the combination of the proposed methods (we show it as a diagonal matrix without duplication of the value for better readability of the plot).

equality of all scores. The scores of easiness to follow the explanation flow are 3.63, 3.83, and 2.93 for *baseline*, *fae*, and *incremental* correspondingly. The difference between *baseline* and *fae* turned out to be not statistically significant according to the T-test (p-value is 0.47, so we cannot reject the null hypothesis). This equal score is understandable because the *baseline* method does not actually deliver any clarification and is pretty short. In terms of understandability, we can see that our method scored higher than the *baseline* and *incremental* methods (the scores are 3.24, 4.1, and 3.15). The results of the T-test let us reject the equality hypothesis between our method and *baseline* and *incremental* as far as p-values are 6E-4 and 1E-5, respectively, so the difference between these scores is statistically significant.

The "Direct comparison" plot shows the answers to the question about the explanation method, which helped to better understand the reasoning process. The final counts are 9, 27, and 5 for *baseline*, *fae*, and *incremental* respectively. Our method clearly outperforms alternative ones. We verify the significance with a binomial test. The null hypothesis of this test is that the vote for *fae* has equal chances of being selected by the participants compared to two other methods (namely that its probability is 1/3). The binomial test yields a p-value of 2E-5, which allows us to reject the null hypothesis.

Finally, the "Combination of methods" plot shows the answers to the question about the potential use of the combination of the proposed methods. We can see that the combination of *fae* and *baseline* received the most votes (19 of 36). This seems reasonable because the baseline answers the simple question "What has changed, given the evidence?" in a straightforward way and shows the dynamic of beliefs about all nodes of the BN, but does not provide any explanation. Whereas our approach explains the probability updates in a more detailed, step-by-step way.

We also briefly analyzed the textual feedback from the participants. Some participants indicated that our algorithm delivers too many details of the reasoning, which may sometimes make perception more difficult and is not always necessary. However, at the same time, other participants found these details useful. The incremental explanation algorithm turned out to be less understandable, partially because some participants failed to understand why some evidence

was disregarded and also because the notion of factors that do not support the probability update is not clear.

Finally, most participants found the visual tips very helpful for understanding the explanation, which corresponded to our initial belief. Indeed, whereas the textual explanation may deliver important information about a BN's reasoning process, it could be particularly difficult to follow the explanation if there is no visual interpretation of it. A good idea may be to deliver such an explanation as an animated picture that could start from the initial state of the BN and then explicitly visualize the *FA* flow (for our method) or highlight the described evidential and intermediate nodes (for incremental method).

The evaluation setup has certain limitations. First, the evaluation is performed on a basic BN of comparatively small size. Second, most participants in this evaluation possess significant knowledge in various fields of computer science or engineering, so their perception of the explanation may be biased. Further steps in the field of BN reasoning explanation have to be taken toward evaluation studies by expert users in the application domain (e.g., healthcare) with real-life BNs that are normally bigger than the one used in our study. However, such an evaluation study is worth a standalone paper [36], so we leave this task for future work.

6. Conclusions

In this work, we introduce a novel approach to the textual explanation of BN inference, which is based on the notion of a factor argument—abstract representations of the flow of information that tangle observations with variables of interest. We present how to use this formalism for the content determination stage of natural-language explanations of approximate reasoning in BNs and, in particular, to decide when to present considerations jointly or separately. We experimentally show that our proposed approach accurately approximates the message-passing algorithm. Finally, we perform a human-driven evaluation using the medical domain BN of the proposed natural-language explanations by comparing it with another explanation method and with the baseline description of the BN. The results of the evaluation show that our method is significantly more understandable than the compared method.

7. Data statement

The BNs used for the experiments in this work are available on bnlearn website [33].

Acknowledgments

This research was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 860621, and the Galician Ministry of Culture, Education, Professional Training, and University and the European Regional Development Fund (ERDF/FEDER program) under grants ED431C2018/29 and ED431G2019/04.

References

[1] D. Doran, S. Schulz, T. R. Besold, What does explainable AI really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).

[2] H. Hagras, Toward human-understandable, explainable AI, Computer 51 (2018) 28–36.

[3] L. Floridi, Establishing the rules for building trustworthy AI, Nature Machine Intelligence 1 (2019) 261–262.

[4] S. Mascaro, Y. Wu, O. Woodberry, E. P. Nyberg, R. Pearson, J. A. Ramsay, A. Mace, D. Foley, T. Snelling, A. E. Nicholson, Modeling COVID-19 disease processes by remote elicitation of causal Bayesian networks from medical experts, medRxiv (2022).

[5] C. S. Vlek, H. Prakken, S. Renooij, B. Verheij, A method for explaining Bayesian Networks for legal evidence with scenarios, Artificial Intelligence and Law 24 (2016) 285–324.

[6] G. Sottocornola, S. Baric, M. Nocker, F. Stella, M. Zanker, DSSApple: A hybrid expert system for the diagnosis of post-harvest diseases of apple, Smart Agricultural Technology 3 (2023) 100070.

[7] E. Kyrimi, S. Mossadegh, N. Tai, W. Marsh, An incremental explanation of inference in Bayesian Networks for increasing model trustworthiness and supporting clinical decision making, Artificial intelligence in medicine 103 (2020) 101812.

[8] J. Keppens, Explainable Bayesian network query results via natural language generation systems, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, 2019, pp. 42–51.

[9] J. Keppens, Explaining Bayesian Belief Revision for Legal Applications., in: JURIX, 2016, pp. 63–72.

[10] E. Reiter, R. Dale, Document Planning, Studies in Natural Language Processing, Cambridge University Press, 2000. doi:10.1017/CBO9780511519857.005.

[11] C. Lacave, F. J. Díez, A review of explanation methods for bayesian networks, The Knowledge Engineering Review 17 (2002) 107–127.

[12] C. Hennessy, A. Bugarín, E. Reiter, Explaining Bayesian Networks in Natural Language: State of the Art and Challenges, in: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, Association for Computational Linguistics, Dublin, Ireland, 2020, pp. 28–33. URL: https://aclanthology.org/2020.nl4xai-1.7.

[13] M. Henrion, M. J. Druzdzel, Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning, in: Uncertainty in Artificial Intelligence, volume 6, 1990, pp. 17–32.

[14] I. Zukerman, K. Korb, R. McConachy, Perambulations on the way to an architecture for a nice argument generator, in: Notes of the ECAI-96 Workshop on Gaps and Bridges:"New Directions in Planning and Natural Language Generation, 1996, pp. 31–36. URL: https://dl.acm.org/doi/abs/10.5555/295240.295901.

[15] I. Zukerman, R. McConachy, K. B. Korb, Bayesian reasoning in an abductive mechanism for argument generation and analysis, in: AAAI/IAAI, 1998, pp. 833–838.

[16] C. Lacave, M. Luque, F. J. Diez, Explanation of Bayesian Networks and influence diagrams in Elvira, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 37 (2007) 952–965.

[17] M. Pereira-Fariña, A. Bugarín, Content determination for natural language descriptions

of predictive Bayesian Networks, in: 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019), Atlantis Press, 2019, pp. 784–791.
[18] A. de Waal, J. W. Joubert, Explainable Bayesian Networks applied to transport vulnerability, Expert Systems with Applications 209 (2022) 118348. URL: https://www.sciencedirect.com/science/article/pii/S0957417422014671. doi:https://doi.org/10.1016/j.eswa.2022.118348.
[19] T. Koopman, S. Renooij, Persuasive contrastive explanations for Bayesian Networks, in: European Conference on Symbolic and Quantitative Approaches with Uncertainty, Springer, 2021, pp. 229–242.
[20] S. Renooij, Relevance for robust bayesian network MAP-explanations, in: International Conference on Probabilistic Graphical Models, PMLR, 2022, pp. 13–24.
[21] H. J. Suermondt, Explanation in Bayesian Belief Networks, Stanford University, 1992.
[22] P. Haddawy, J. Jacobson, C. E. Kahn Jr, BANTER: a Bayesian network tutoring shell, Artificial Intelligence in Medicine 10 (1997) 177–200.
[23] G. A. Vreeswijk, Argumentation in Bayesian belief networks, in: Argumentation in Multi-Agent Systems: First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers 1, Springer, 2005, pp. 111–129.
[24] J. Keppens, Argument diagram extraction from evidential Bayesian Networks, Artificial Intelligence and Law 20 (2012) 109–143.
[25] G.-E. Yap, A.-H. Tan, H.-H. Pang, Explaining inferences in bayesian networks, Applied Intelligence 29 (2008) 263–278.
[26] S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, A two-phase method for extracting explanatory arguments from Bayesian Networks, International Journal of Approximate Reasoning 80 (2017) 475–494. URL: https://www.sciencedirect.com/science/article/pii/S0888613X16301402. doi:https://doi.org/10.1016/j.ijar.2016.09.002.
[27] M. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, San Diego, California, 2016, pp. 97–101. URL: https://aclanthology.org/N16-3020. doi:10.18653/v1/N16-3020.
[28] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[29] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[30] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society. Series B (Methodological) 50 (1988) 157–224. URL: http://www.jstor.org/stable/2345762.
[31] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.
[32] D. Madigan, K. Mosurski, R. G. Almond, Graphical Explanation in Belief Networks, Journal of Computational and Graphical Statistics 6 (1997) 160–181. doi:10.1080/10618600.1997.10474735.
[33] bnlearn, 2024. URL: https://www.bnlearn.com/bnrepository/.
[34] A. Darwiche, Modeling and reasoning with Bayesian networks, Cambridge university

press, 2009.
[35] Netica, 2024. URL: https://www.norsys.com/netica.html.
[36] R. Butz, R. Schulz, A. Hommersom, M. van Eekelen, Investigating the understandability of XAI methods for enhanced user experience: When bayesian network users became detectives, Artificial Intelligence in Medicine 134 (2022) 102438. URL: https://www.sciencedirect.com/science/article/pii/S0933365722001907. doi:https://doi.org/10.1016/j.artmed.2022.102438.
[37] Bayes Sever, 2024. URL: https://www.bayesserver.com/examples/.

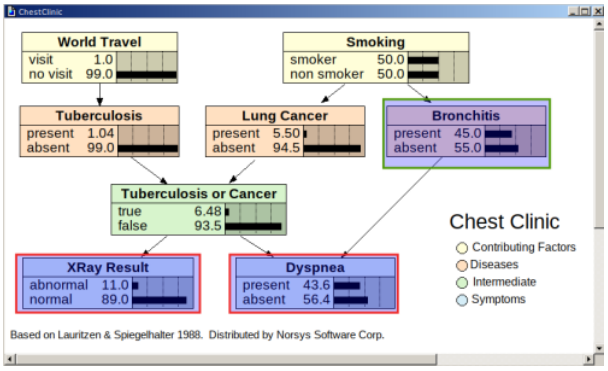
A. Human-driven evaluation survey

Textual explanation:

The patient has abnormal XRay results and does not have dyspnea, so the probability of the absence of Bronchitis has increased from 55.0% to 80.7%.

Visual explanation:

Before evidence



After evidence

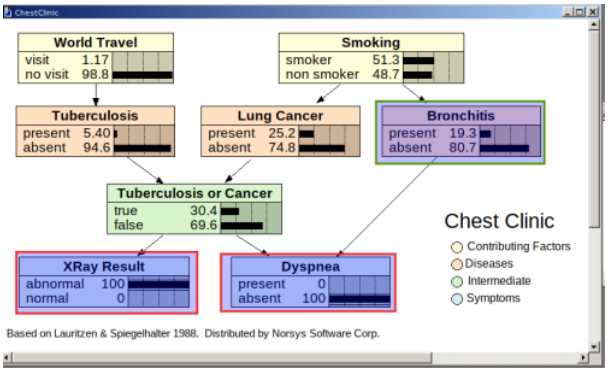


Figure 10: Example of baseline explanation in the interface of the human-driven evaluation survey

Target node <Bronchitis> and factors affecting it

The likelihood of <Bronchitis> = <absent> is 80.7%
This is 25.7% INCREASED compared to the general case (without any evidence)

The significant factors that support the INCREASED likelihood of <Bronchitis> = <absent>:
Dyspnea = absent

There are no significant factors that do not support the INCREASED likelihood of <Bronchitis> = <absent>

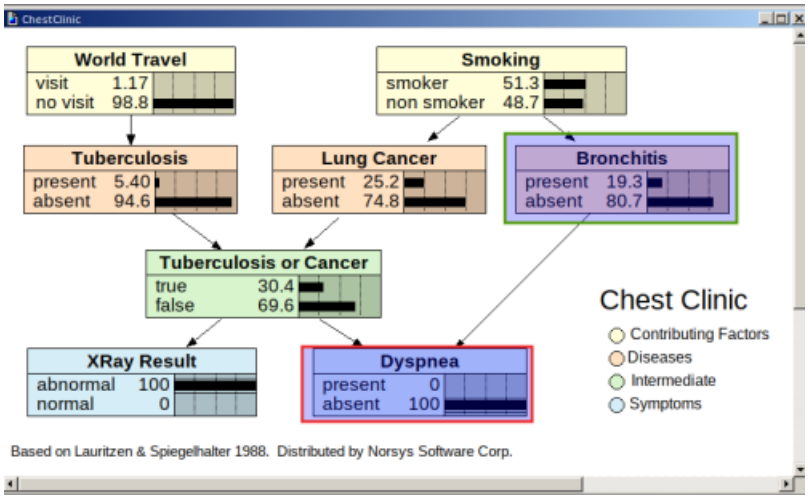


Figure 11: Example of incremental explanation (first explanation stage described in [7]) in the interface of the human-driven evaluation survey

Important elements for predicting <Bronchitis> are:

1. <Tuberculosis_or_Cancer>

The likelihood of <Tuberculosis_or_Cancer> = <false> is 69.6%
This is 23.9% DECREASED compared to the general case (without any evidence)

There are no significant factors that support the DECREASED likelihood of <Tuberculosis_or_Cancer> = <false>

The significant factors that do not support the DECREASED likelihood of <Tuberculosis_or_Cancer> = <false>:
Dyspnea = absent

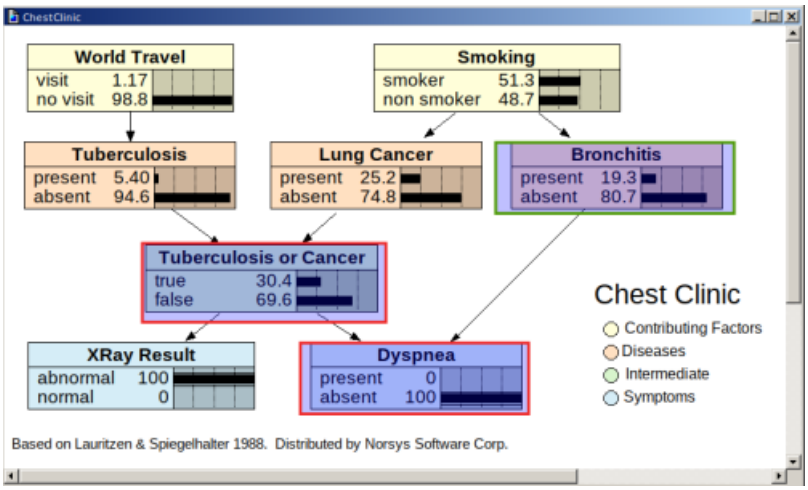


Figure 12: Example of incremental explanation (second and third explanation stages described in [7]) in the interface of the human-driven evaluation survey

Overall, based on the evidence provided, the probability of <Bronchitis> = <absent> INCREASED from 55.0% to 80.7%

Below you will see the paths of update of the probability.
They are sorted in descending order by the effect to the change of the probability of the target node <Bronchitis>

Argument Path #1 : From evidence <XRay_Result>,<Dyspnea> to target <Bronchitis>

We have observed that <XRay_Result> is <abnormal> and <Dyspnea> is <absent>.

The updated probability of <XRay_Result> = <abnormal> is evidence that the intermediate node <Tuberculosis_or_Cancer> becomes strongly more likely to be <true>

The updated probability of <Tuberculosis_or_Cancer> = <true> and <Dyspnea> = <absent> is evidence that the target node <Bronchitis> becomes strongly more likely to be <absent>

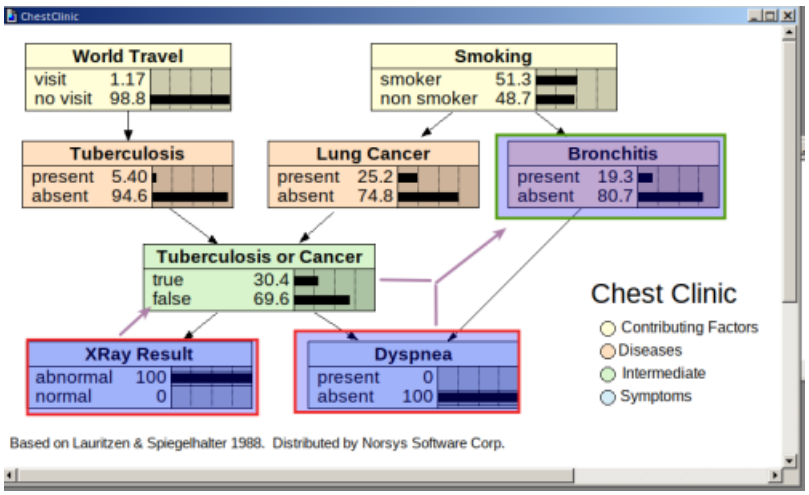


Figure 13: Example of our explanation method in the interface of the human-driven evaluation survey

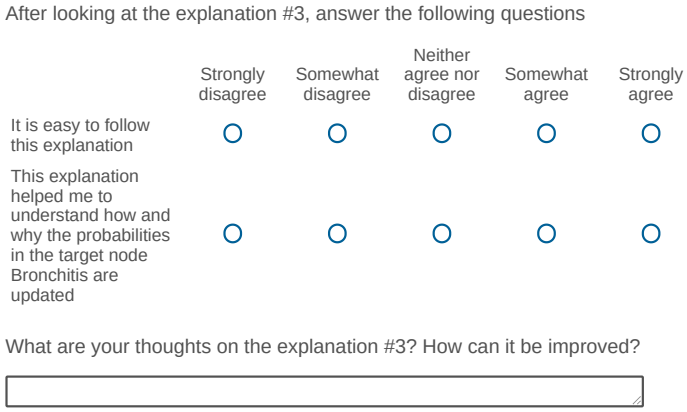


Figure 14: Example questions about definite explanation method in the interface of the human-driven evaluation survey

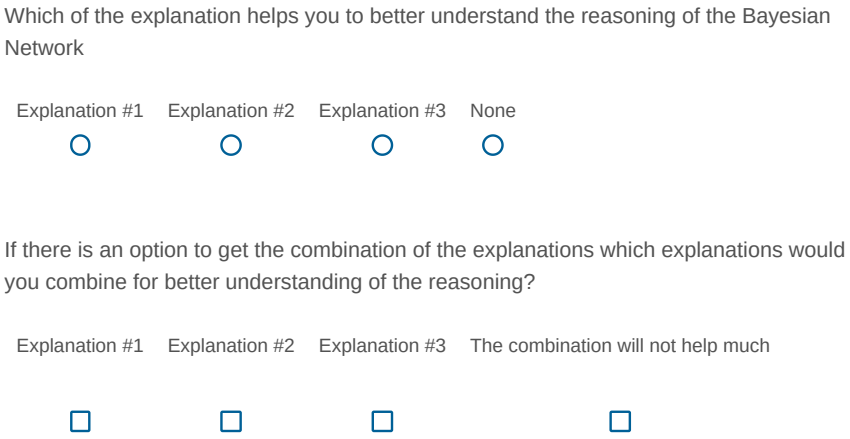


Figure 15: Example questions about all explanation methods within one case in the interface of the human-driven evaluation survey