Computing contrastive, counterfactual explanations for Bayesian networks

Tara Koopman

July 16, 2020

Student number: 5713811
Supervisor: Silja Renooij
Program: Artificial Intelligence, Utrecht University

Abstract

In recent years it is explored how counterfactual statements can be used to explain the value of a specific target variable of several Artificial Intelligence systems. However, how counterfactual statements can be used to explain a Bayesian Network (BN) is a relatively unexplored topic. Because people generally prefer an explanation where one value is contrasted against another, we want to give an explanation that is both contrastive and counterfactual to explain a certain target variable in a BN. After giving a definition for a contrastive, counterfactual explanation and giving a naive approach of computing all explanations, we first constructed an algorithm to more efficiently find all explanations in an enhanced subset lattice for evidence variables that are binary valued. Secondly we explored how a monotonicity relation between the evidence and the target can be exploited to more efficiently compute all explanations for evidence that is not binary valued. We were able to derive several propositions about the inclusion of evidence variables in an explanation based on the respective ordering of the observed value for the evidence variable and the most probable and expected values for the target. With these propositions we constructed an algorithm that finds all explanations with a breadth-first search through the monotonicity enhanced subset lattice. We concluded our research by providing two different methods of selecting those explanations that are most useful for a user and giving different templates of how the explanations can be presented to the user in a textual way.

1 Introduction

Since the introduction of decision support systems, it is researched how the results of these systems should be presented to the users, such that they could benefit from them the most. In early studies it became clear that the system's ability to explain its decisions is considered its most important feature by the users (Teach & Shortliffe, 1981).

Bayesian networks (Jensen & Nielsen, 2007) can be used as a decision support system. A lot of research has been done on how these networks can be explained which resulted in a wide range of different explanation methods. There are three different aspects of a Bayesian network that can be explained; the model itself, how unobserved variables relate to evidence in the model, and the reasoning process that led to the most probable value of a target variable in the model (Lacave & Díez, 2002). This last type of explanation is researched the most. The aim of explaining the reasoning of the model is to give a clear explanation of the inference process that resulted in the most probable value for a target variable.

In an evaluation study, Suermondt & Cooper (1993) measured the effect on user confidence and accuracy when an explanation of the inference was given to the user in addition to the decision of a decision support system based on a Bayesian network. The domain of the network was anesthesia and the explanations were generated by INSITE, a program developed by Suermondt. One result of the evaluation was that users who did not get an explanation grew less confident when they disagreed with the assessment of the system, while this effect was not present with users who were provided with an explanation. Another conclusion was that explanations were used to resolve conflicts between the preferred decision of the system and the user, which made the user avoid incorrect decisions. This led to an improved diagnostic accuracy of the user (Suermondt & Cooper, 1993). From this last result we can conclude that explanations are especially useful when a user initially disagrees with a system, because in those cases providing an explanation can lead to a better diagnostic accuracy. When agreeing with the system, an explanation would merely confirm the user's ideas and reformulate a similar reasoning as the user has used to reach the same conclusion. If the user disagrees with the decision, the explanation of the reasoning of the system could give the user new insights in the importance and interactions among different variables and could point the user to things he might initially have overlooked.

People naturally ask for contrastive explanations, where an event or outcome that did occur is explained in relation to another event or outcome that did not occur (Miller, 2019). Especially when unexpected behaviour is encountered, users tend to ask a *Why not*-question (Lim & Dey, 2009). In other words, people are interested in an explanation why a certain decision was taken by a system instead of the decision they were expecting. In this situation there are two different cases; either the decision suggested by the system or the decision expected by the user is the better choice. In the first case can a contrastive explanation point the user to the flaws in his own reasoning. In the second case can the user use the explanation to determine what led to the suboptimal result of the system. In both cases will the explanation probably lead to a better cooperation between the user and the system, which can further increase the trust of the user in the system.

This thesis presents a research that contributes to the methods that are used to explain the the value of a target variable of a Bayesian network (BN). In Section 2 we first give a short introduction to Bayesian networks. We present the results of a literature research in Section 3. This literature research consists of two parts. In the first part we define what an explanation is and why it is needed for Bayesian networks and systems using Artificial Intelligence (AI) in general. From this we conclude that people generally prefer explanations that are both contrastive and local, where a local explanation for an AI system explains one specific action or outcome of the system. The second part of the literature research gives an overview of the existing explanation methods that are currently used for BNs and AI systems. From this overview we conclude that none of these methods use contrastive explanations. We also conclude that in recent years counterfactuals are used to explain AI systems. A counterfactual is a statement of how changing the value for certain evidence variables can make a different value the most probable for a target. It is not yet explored how counterfactuals can be used to explain BNs. We combine the contrastive with the counterfactual explanation to provide a local explanation for the reasoning of a BN to a user. This results in the following research question:

> **How can a contrastive, counterfactual explanation for a target variable of a Bayesian network be obtained and conveyed to the user?**

To answer the research question we divide it in three different parts. The first part of the research focuses on how a contrastive, counterfactual explanation can be defined. The result is given in Section 4.

The second part of the research focuses on the question of how the explanation as defined in Section 4 can be computed. This will be the largest part of the research. In Section 5 we start with a naive approach of computing

all explanations that directly follows from the definition. We will see that this approach is inefficient and does redundant work. In Section 6 we explore how a subset lattice can be used to more efficiently find all explanations for evidence that is binary valued. We find that this method can be extended to also work for evidence that is not binary valued, however it will then have a too large complexity to work in practice. We conclude that we need some additional information about the evidence to be able to derive all explanations more efficiently. In Section 7 we give two different definitions of monotonicity and prove several propositions about the inclusion of an evidence variable in an explanation based on the ordering of its observed value. In Section 8 we use these findings to construct an algorithm that finds all explanations in the lattice for evidence that has a monotone relation with the target.

The last part of the research explores how the explanation can be given to the user. This is presented in Section 9 along with some examples of how an explanation can be further adjusted to the needs and preferences of a user. In Section 10 a conclusion is given about the results presented in this thesis. At last we give some entry points for additional research in Section 11.

2 Bayesian Networks

In this section we give a short introduction to Bayesian networks. In Section 2.1 we give a definition of a Bayesian network and introduce some notation that will be used throughout this thesis. Afterward we give an example of a Bayesian network in Section 2.2.

2.1 Definition

A Bayesian network (BN) represents a joint probability distribution on a set of random variables (Jensen & Nielsen, 2007). These variables can take on values that are mutually exclusive and collectively exhaustive. First we introduce some notation for variables, sets of variables and their value configurations. A capital V_i denotes a single variable, a lower case v_i is used to denote a specific value for V_i. A bold, capital $\mathbf{V}$ is used to denote a set of variables, with bold, lowercase $\mathbf{v}$ a value configuration for $\mathbf{V}$. If not further specified $\mathbf{v}$ is the value configuration where all variables in $\mathbf{V}$ take on their observed values. Given two value configurations $\mathbf{v}$ and $\mathbf{w}$ for $\mathbf{V}$ and $\mathbf{W}$ respectively, $\mathbf{v}$ and $\mathbf{w}$ are consistent if for all V_i with $V_i \in \mathbf{V}$ and $V_i \in \mathbf{W}$, V_i takes on the same value in $\mathbf{v}$ as in $\mathbf{w}$. Otherwise, $\mathbf{v}$ and $\mathbf{w}$ are inconsistent.

A BN consists of two parts; a qualitative and a quantitative part. The qualitative part is a directed, acyclic graph. The nodes in this graph correspond to the random variables of the joint probability distribution. If a node V_j has an arrow to node V_i in a graph, V_j is called the parent of V_i. $\pi(V_i)$ denotes the set of all parents of V_i. Given an arrow from V_i to V_j, V_j is called the child of V_i. The descendants of V_i are recursively defined as the children of V_i and all descendants of the children of V_i. The Markov blanket of a variable V_i is a special set of variables related to V_i. This is defined in the following way.

Definition 2.1. (Markov blanket) Given a directed, acyclic graph G, the Markov blanket of a variable V_i is a set consisting of all parents of V_i, all children of V_i and all parents of the children of V_i in G.

The graph of a BN represents the independences among the variables. Given a set of evidence, it is directly derived from the structure of the graph if two variables are independent given this evidence. The concept of d-separation is used to derive if two variables are independent. To be able to define d-separation, we first give a definition for blocked chains in a directed, acyclic graph. A chain in a directed graph G is defined as a path on the graph that is constructed by replacing each arc in G by an edge.

Definition 2.2. (Blocked chain) Given a directed, acyclic graph G with a set $\mathbf{E}$, a chain c from V_i to V_j is blocked if there is a variable W_i on c such that one of the following statements hold:

- W_i has at most one incoming arc on c and $W_i \in \mathbf{E}$.
- W_i has two incoming arcs on c and $W_i \notin \mathbf{E}$ and for all descendants D_i of W_i we have $D_i \notin \mathbf{E}$.

If a chain is not blocked, it is called an *active* chain. Based on the notion of blocked chains, we define d-separation in the following way.

Definition 2.3. (D-separation) Given a directed, acyclic graph G with a set $\mathbf{E}$, V_i and V_j are d-separated by $\mathbf{E}$ if all chains from V_i to V_j in G are blocked by $\mathbf{E}$.

If a variable V_i is d-separated from V_j by $\mathbf{E}$, V_i is independent of V_j given $\mathbf{E}$. This means that changing the probability distribution of V_i will not impact the probability distribution of V_j. Given the Markov blanket of V_i, V_i is d-separated and thus independent of all other variables in the graph.

The quantitative part of the BN consists of conditional probability functions for every node in the graph. A probability function of a variable V_i gives the probabilities for all possible values of V_i given all value configurations of $\pi(V_i)$ in the graph. $P(V_i|v_j)$ denotes the conditional probability distribution of V_j given the value v_j for V_j. $\top(V_i|v_j)$ denotes the most probable value of V_j given the value v_j for V_j. The most probable value for a variable is also called the mode for this variable.

We now define a BN in the following way.

Definition 2.4. (Bayesian network) A Bayesian network is a tuple (G, Γ) where

- $G = (\mathbf{V}, \mathbf{A})$ is a directed, acyclic graph, with $\mathbf{V} = \{V_1, ..., V_n\}$ the nodes and $\mathbf{A}$ the arcs in the graph.
- Γ is a set of real-valued probability functions that specify $P(V_i|\pi(V_i))$ for all $i = 1, ..., n$.

A BN contains all information necessary to compute a joint probability distribution on the variables. Given a BN with variables $\mathbf{V}$ the joint probability distribution $P(\mathbf{V})$ is defined by

$$P(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} P(V_i|\pi(V_i))$$

Computing a joint probability distribution in a BN is called inference. Several inference algorithms have been developed that more efficiently compute these joint probability distributions by exploiting the indepenences that are represented in the graph. An example of such algorithms is the algorithm of Pearl (1988).

In general we are not interested in a full joint probability distribution. Users of a Bayesian network are usually interested in the probability distribution of one or more specific variables in the network. We call these the target variables. In the remainder of this thesis we assume there is always one target variable. We use T to denote this target. Given all variables $\mathbf{V}$, not all variables in $\mathbf{V} \setminus \{T\}$ are observable. We call the set of observable variables the evidence set. In the remainder of this thesis we use $\mathbf{E}$ to denote this set. $\mathbf{e}$ is used to denote the observed value configuration for $\mathbf{E}$. The remaining variables $\mathbf{V} \setminus (\mathbf{E} \cup \{T\})$ in the BN are unobservable. We call this set the hidden variables, this is denoted with $\mathbf{H}$.

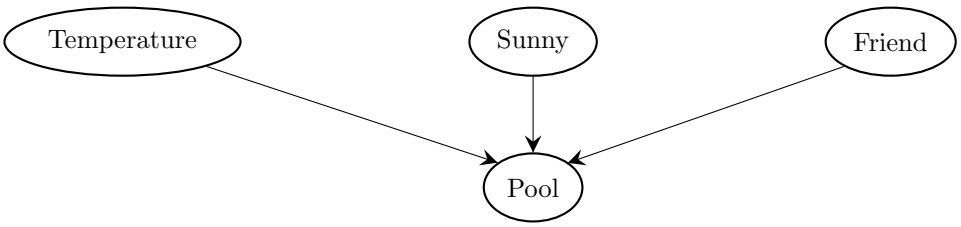
2.2 Swimming pool example

We define the swimming pool example to illustrate the use of a BN. The graph of the BN for this example with its associated probabilities is given in Figure 1.

This example has four different variables. The first variable is $Temperature$, which can take on three different values; it is either colder than 25 degrees, between 25 and 30 degrees or hotter than 30 degrees. $Sunny$ is a binary valued variable that indicates whether it is sunny or not. $Friend$ is also binary valued and states if our friend goes to the pool. The last variable is $Pool$, this binary valued variable indicates if we will enjoy our day at the pool.

This BN is used to decide if we go to a pool on a given day, based on the probability that we enjoy spending a day at the pool. If $true$ is the most probable value for $Pool$, we decide to go to the pool. Otherwise, we decide not to go. It follows that $Pool$ is our target variable. All other variables are observable, so $\mathbf{E} = \{Temperature, Sunny, Friend\}$.

Before any observations are done, the probability that we enjoy our day at the pool is $P(Pool = true) = 0.59$. Because it is more likely that we enjoy a day at the pool, we decide to go to the pool today. We now observe that it is not sunny, our friend is not at the pool and the temperature is below 25 degrees. Given these observations, the probability that we enjoy a day at the pool drops; $P(Pool = true|Temp < 25, Sunny = false, Friend = false) = 0.30$. We now decide to stay at home today.



$P(Temp < 25) = 0.5$ $P(Temp = 25 - 30) = 0.3$ $P(Temp > 30) = 0.2$

$P(Sunny = true) = 0.6$ $P(Sunny = false) = 0.4$

$P(Friend = true) = 0.4$ $P(Friend = false) = 0.6$

$P(Pool = true|Temp < 25, Sunny = false, Friend = false) = 0.3$
$P(Pool = true|Temp < 25, Sunny = false, Friend = true) = 0.35$
$P(Pool = true|Temp < 25, Sunny = true, Friend = false) = 0.4$
$P(Pool = true|Temp < 25, Sunny = true, Friend = true) = 0.7$

$P(Pool = true|Temp = 25 - 30, Sunny = false, Friend = false) = 0.35$
$P(Pool = true|Temp = 25 - 30, Sunny = false, Friend = true) = 0.7$
$P(Pool = true|Temp = 25 - 30, Sunny = true, Friend = false) = 0.75$
$P(Pool = true|Temp = 25 - 30, Sunny = true, Friend = true) = 0.8$

$P(Pool = true|Temp > 30, Sunny = false, Friend = false) = 0.77$
$P(Pool = true|Temp > 30, Sunny = false, Friend = true) = 0.8$
$P(Pool = true|Temp > 30, Sunny = true, Friend = false) = 0.875$
$P(Pool = true|Temp > 30, Sunny = true, Friend = true) = 0.9$

Figure 1: The BN representing the swimming pool example, the numbers in the last table represent the probability of enjoying our day at the pool.

3 Explanation

To be able to construct a method that extracts explanations from a BN, we need to understand what an explanation is and why and when it is needed. We answer different questions about explanations from an AI perspective in Section 3.1.

In Section 3.2 we describe different methods that are currently used to explain AI systems. Some of these methods can be applied to a wide range of AI systems, others can only be applied to neural networks. We describe explanation methods that are specifically designed to explain BNs in Section 3.3. In Section 3.4 we contrast the explanation methods for AI systems against the explanations methods for BNs. We use this to determine if there are some concepts used in the explanations for AI systems that are not yet exploited by explanation methods for BNs.

3.1 What is an explanation?

According to the Oxford Dictionary an explanation is *"a statement or account that makes something clear"* or *"a reason or justification given for an action or belief"*(*Oxford English dictionary*, n.d.). After an extensive search, no formal definition of explanation could be found that is used as a standard in the literature about explanation methods for AI systems. Most studies that research explanation methods do not give a clear definition of what they see as a (good) explanation. Before discussing different explanation methods for AI systems in general and Bayesian networks in particular, we first consider different questions about explanations, which will be answered from an AI perspective. These answers are primarily based on the research of Miller (2019) and Adadi & Berrada (2018).

3.1.1 Why is an explanation needed?

There are four reasons why explanations for AI systems are needed. First of all, an explanation can *justify* the results of the system to a user, who is not familiar with AI technology. Secondly, explanations can help the developers of the system to *control* the system by easily identifying mistakes. Thirdly, developers can more easily *improve* the system, because they understand it better if it is explained. At last, an explanation can help to *discover* new facts, because knowledge learned from the system is transferred to the users and developers of the system with an explanation. In this thesis, we will focus on explanations that try to justify results to the user. The goal of explanations to justify is to help the user understand the system and to gain trust of the user. Without the trust of the user, AI systems will not be easily deployed in real life domains.

There is not the same need for justification for all types of AI systems. There are two factors that influence the need for explainability. The first factor is the domain in which the system operates. In domains where the decision or the action taken by the system has high consequences, the demand for an explanation is higher (Adadi & Berrada, 2018). Consider for example an algorithm used in the entertainment domain that recommends movies to a user. There will be negligible effects of a wrong recommendation, so there is not a high demand for explanations. However, if an AI system is used in a medical domain to diagnose a patient, a mistake made by the system could have severe consequences. In this situation an explanation is needed to possibly prevent these mistakes and to gain the trust of the user of the system.

The second factor is the transparency of the system. Some AI methods, for example decision trees, are considered to be interpretable. For these methods an explanation is not necessary. Other systems, for example neural networks, are considered black-box models, because the processes that transform the input to the output are not transparent. For such systems there is a high demand for explanations to get some insights in how the system works.

Bayesian networks do not fall in either of these categories. They are not black-box models, because the inference algorithms and the independences among the variables in the network are transparent and understandable for BN experts. However, they are not easily understood for people without knowledge of these systems. Because the explanations to justify are used to gain the trust of the user of the system, BNs could also benefit from explanations.

3.1.2 At whom is the explanation aimed?

An explanation is a social interaction between the explainer, someone giving the explanation, and the explainee, someone receiving the explanation (Miller, 2019). For the remainder of this thesis, we will only consider cases where the explainer is an AI system, that is operating in a certain domain.

After evaluating different explanation methods for AI systems, we concluded that there are three different types of explainees. First of all, the explainee can be an expert on a certain AI system. An AI expert will mainly use the

explanations to control or to improve the system. An AI expert is in some sense the explainer and the explainee at the same time. He can develop the methods that give the explanation, but he can also benefit from the explanation by using it as a way to debug and improve the system. An explanation given to an AI expert, can contain technical information and can be very specific about the used algorithms and methods.

Secondly, the explainee can be someone who is an expert in the domain in which the AI system operates, but is not familiar with AI methods. For example, when considering an AI system that diagnoses patients in a medical domain, the explainee is a doctor who has vast knowledge about a certain disease. A domain expert mainly uses the explanation to justify the results of the system and to resolve disagreement when the results of the system do not align with the expected result from the domain expert (Gregor & Benbasat, 1999). An explanation aimed at a domain expert can not contain technical details about the system, but it can contain terminology specific to that domain.

The last type of explainee is the subject of the system who has neither domain knowledge nor knowledge about AI. Consider for example a financial domain where it is decided based on an AI system if someone gets a loan. The person for who this is decided is the subject of the system. A subject uses an explanation to justify the results of the system. According to European law, subjects have the right of explanation to understand decisions made about them based on AI systems and to give a ground to contest that decision if they do not agree with it (Wachter et al., 2017). An explanation aimed at a subject, should be easily understandable for anybody without domain knowledge or knowledge about AI.

3.1.3 What is explained?

Explanation methods can be divided into global or local explanations (Adadi & Berrada, 2018). Global explanations try to explain the whole logic and all possible outcomes of the entire model. Local explanations try to justify a specific outcome or action of the system. When using an AI as a decision support system, giving local explanations for multiple decisions gives a better decision making accuracy of the user than giving a single global explanation of the system (Gregor & Benbasat, 1999). In addition a global explanation of a complex model is hard to achieve in practice (Adadi & Berrada, 2018). So most explanation methods focus on local over global explanations.

Lacave & Díez (2002) define three categories of explanations when considering Bayesian networks. The first category is the explanation of the model, this corresponds to a global explanation. The second category is the explanation of the reasoning. With this kind of explanation the value for a target variable of the network and the reasoning process that led to this value is justified, so this corresponds to a local explanation. The third category is the explanation of the evidence. The available evidence is explained by giving the most probable configuration of unobserved variables. Such an explanation can be considered local, because it considers one specific case, namely one specific set of evidence variables. However, it does not completely fall into this category, because no result of the system is explained. In this way it can also be considered a global explanation, because based on one specific case the whole system is explained.

3.1.4 How is the explanation given?

The explanation can be presented to the user in different ways. It can be given in a textual way. For example by listing the most influential variables, giving if-then rules based on the model or by stating what variables can be changed to result in a different value for the target. In case of a BN, a textual representation does not only consist of words, but it can also present the probabilities in the graph (Lacave & Díez, 2002).

An explanation can also be given in a visual way. This can for example be done by plotting the probability of the target variable against the value of an evidence variable in a system. In case of an explanation for an BN, a visual explanation can consist of showing the graph of the BN and using colors and thickness of graph arcs to indicate the influences among variables (Lacave & Díez, 2002). Textual and visual explanations can also be combined.

An explanation can be given in an interactive or non-interactive way. With an interactive explanation, a user can ask for clarification about concepts he does not yet understand or ask for more details. In a non-interactive way, the explanation is given without any user input.

3.1.5 What is a good explanation?

An explanation given by an AI system should be easily understandable for the users of the system. To be understandable for the domain expert or the subject of the system, they should resemble explanations that people give each other (Miller, 2019). After surveying multiple publications on explanations in the field of psychology and philosophy, Miller (2019) gives the following conclusions about what elements of human explanations should

be used in explanations of AI systems. First of all, an effective explanation should consist of giving causes for a result of the system. Making generalisations on statistics and probabilities is considered unsatisfying by users of a system. Secondly, users prefer simple explanations, where fewer causes are given for a result, over more complex explanations where more reasons are given to explain the same result. Another conclusion is that people mostly use contrastive explanations. These are explanations where somebody does not ask why a certain result is obtained, but rather why a certain result is obtained instead of some other result. Finally, an explanation should always be a social interaction between the user and the system, where the knowledge of the user is taken into account. An explanation can be given in a static way, where the explanation stays the same every time the system is used, or in a dynamic way, where the explanation is adjusted as the user learns more about the system (Lacave & Díez, 2002). According to the last conclusion of Miller there is a preference of the dynamic over the static way.

3.2 Existing explanation methods for AI systems

Adadi & Berrada (2018) give an introduction to the rise of explainable AI and give an overview of some explanation methods for AI systems. The overview of different explanation methods for AI systems we give in this section is based on this review. Because of the current interest in neural networks, most of these methods are specifically designed to explain neural networks. However, some methods can be applied to a wider range of algorithms and systems.

We give a concise introduction to neural networks to better understand the explanation methods given below. A neural network uses multiple input variables to compute the value of one or more output variables. A neural network is composed of several nodes, which all have a so-called activation function (Russell & Norvig, 2016). The nodes are arranged in different layers. Each layer is connected to the next layer by links between the nodes, where each link has its own weight. The first layer is called the input layer; the nodes in this layer correspond to input variables of the network. The nodes in all layers compute their output based on the weighted input they receive and their activation function and then send this output through the links to the nodes in the consecutive layer. The last layer is called the output layer. The node or nodes in this layer correspond to the output variables in the network. All layers between the input and the output layer are called the hidden layers. In Section 2 we defined a set of hidden nodes in a BN. Note that the meaning of *hidden* is different in both cases. In the Bayesian network, the hidden nodes are variables whose value can not be observed, but whose meaning is well defined. For neural networks, the nodes in the hidden layers do not represent a real world variable.

The weights of the links in a neural network and the activation functions of the nodes are learned based on examples in a dataset. For all explanation methods discussed below it is assumed there is a dataset that is used to train the model that is explained. Some methods specifically use this dataset to generate an explanation.

3.2.1 Using variables to explain

Partial dependence plots (PDP) and individual conditional expectation (ICE) plots are methods that try to give insight in the relation between one specific input variable and the output variable. ICE plots the estimated conditional expectation curves, which is a plot of the output variable against a variable of interest (Goldstein et al., 2015). An ICE plot is generated in the following way. The value of a variable of interest is incrementally updated over its whole range of different values and the expected output is computed with the model given the value for all other variables of one specific instance in a dataset. These expected output values are now plotted against the variable of interest. This procedure is repeated for every instance in the dataset, resulting in a plot with as many lines as instances in the dataset.

ICE is based on PDP. Where ICE plots one line for every instance in the dataset, PDP shows just one line. PDP plots the average of the output variable against a variable of interest (Friedman, 2001). This result in a curve that is the average of all curves given by ICE. This method of explanation is used by researchers in different fields, such as ecology and criminal justice, to gain better understanding of the model they use (Adadi & Berrada, 2018).

Both ICE and PDP are useful in showing how one variable influences the output, but they do not tell how variables interact with each other or if some variables are more important than others. How important one variable is to the value of a the output variable of a model can be computed with the Shapley values (Lipovetsky & Conklin, 2001). The Shapley values are computed with the contribution of one variable. Given a set of input variables, the contribution of one variable is defined as the difference between the output value of the model if the value of this variable is known and the expected output value if it is unknown. However, the contribution of a variable does not take the interactions among variables into account. To also give some indication of these interactions, the contribution of a subset of variables is computed. The contribution of a subset of variables is the difference in the output value if the values of all variables in this subset are known and if no value of any variable is known.

Given a set of evidence $\mathbf{E}$, the Shapley value of variable $E_i \in \mathbf{E}$ is computed by summing over the difference in the contribution of $\mathbf{E}'$ and $\mathbf{E}' \cup \{E_i\}$, for all possible subsets $\mathbf{E}' \subseteq \mathbf{E} \backslash \{E_i\}$ and normalizing by the amount of interactions in the subsets.

Computing a Shapley value has an exponential time complexity, so it is not feasible to practically use them. Štrumbelj & Kononenko (2014) defined a way of approximating the Shapley values and proved in an evaluation study that users have an improved understanding of a model if an explanation in the form of variable contributions was given.

3.2.2 Rule extraction

A way of globally explaining AI systems is with the use of rule extraction. These kind of methods try to formulate if-then rules based on a neural network. There are three different approaches of rule extraction.

The first approach is the pedagogical approach. This approach only uses the input variables and the output variable to generate rules. Because only the input and the output layers are considered, the pedagogical approach can be used on complex neural networks with multiple hidden layers. An example of a method that uses the pedagogical approach is OSRE. Where previous methods only worked on binary valued variables, OSRE is developed to also work on ordinal data and it can easily be extended to be used with quantised continuous data (Etchells & Lisboa, 2006). The proposed algorithm delivers rules in the form of $X_i = x_i \wedge ... \wedge X_n = x_n \rightarrow a > 0.5$, where $\{X_1, ..., X_n\}$ is a subset of the input variables, x_i is a subset of the possible values of X_i and a is the output variable. These rules are refined by deleting unnecessary conjuncts and ordered by their specificity value. Rules with a specificity below a certain value are removed. OSRE was tested on different datasets and delivered understandable rules which were able to classify over 90% of the used dataset correctly.

Another approach is the decompositional approach. This approach focuses on extracting rules on the level of individual nodes in the network (Adadi & Berrada, 2018). DeepRED used this approach on deep neural networks (Zilke et al., 2016). The rules that are constructed by this algorithm are in the same form as OSRE, but they are constructed with the help of the hidden nodes. Working from the output layer back to the input layer, a decision tree is made for each layer based on how the nodes in the previous layer are connected to nodes in the current layer. From these trees intermediate if-then rules are computed, that tell how the current layers is connected to the previous. The consecutive rules are then combined in such a way that the result are rules where the if statements are variables from the input layer and the then statements are variables in the output layer.

The last approach is the eclectic approach. This approach combines elements from the above two approaches to come up with rules. FERNN is an example of such an approach (Setiono & Leow, 2000). First FERNN identifies the most relevant hidden nodes and builds a decision tree based on these nodes. The next step is to remove irrelevant connections from the input layer to the first hidden layer. Input connections are seen as irrelevant when removing this connection does not affect the classification accuracy (Setiono & Leow, 2000). Based on the remaining input connections the decision tree is transformed to only include these connections. The last step is to generate if-then rules by replacing each node splitting condition in the decision tree by a symbolic rule.

3.2.3 Surrogate models

Some simple models, such as linear regression models and decision trees with small depth, are easily interpretable. The idea behind a surrogate model is to transform a complex model into an interpretable model as a way of explanation. A drawback of this approach is that there is no guarantee that the surrogate model is representative of the complex model (Adadi & Berrada, 2018).

LIME is a method that transforms all types of classifiers and regressors into interpretable surrogate models (Ribeiro et al., 2016). These models are used to explain a single observation. The surrogate model does not necessarily have to be the same type of model as the original model. For example if the original model is a random forest, the surrogate model can be a decision tree with a small depth. If the original model is a neural network that is used for image classification, the surrogate model can be a neural network that only uses the subset of pixels that was most important for the classification of the image of interest.

The first step that LIME uses to construct a surrogate model is to generate a new dataset by randomly permuting samples of the original dataset and computing the expected target values for the permuted samples. The instances of this new dataset are weighted based on the proximity to the instance of interest. Now weighted surrogate models are trained based on these weighted samples. The surrogate model that is used for explanation is the model with the lowest model complexity and the lowest loss in predicting the instance of interest. How the loss, model complexity and proximity are defined depends on the type of surrogate model that is used.

3.2.4 Counterfactuals

Wachter et al. (2017) propose the idea to use counterfactual statements as a local explanation for the target of a system. Given an observation, a counterfactual is a statement of what elements of that observation should be different to result in a different or desired value for the target. For example if a student with an average grade of 6.5 is not accepted to a course, a counterfactual statement is; *If the average grade of the student had been* 7.0 *or higher, he would have been accepted to the course.*

If a counterfactual is used as an explanation it should have minimal changes to the actual observation and those changes should be reasonable in the real world. Wachter et al. argue that counterfactual explanations are better interpretable than other explanation methods, because no knowledge about complex machine learning systems is required. This makes the explanation not only understandable for the developers of the system but also for the users of the system. In addition are counterfactuals easily computable. Wachter et al. give the following computation for finding the best counterfactual:

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

Where x_i is the original instance, x' is the counterfactual statement, y' is the desired target value, $d()$ is a distance function that measures how far away from each other x_i and x' are and λ is a weight function. λ is used to find a balance between counterfactual statements that have the smallest changes to x and counterfactual statements whose target values are closest to y'. Small values for λ favor the first kind and larger values prefer the latter. The hardest part of computing counterfactuals is to define a useful distance measure and weight function.

Grath et al. (2018) expand on this idea of counterfactual explanations in the domain of credit applications. They note that not all variables in a problem are equally important and that a counterfactual explanation containing fewer variables is better interpretable for the user. To compute counterfactuals based on these observations, Grath et al. introduce a weight vector, called the global feature importance, to the distance function. The global feature importance is the variance between the actual value of a variable and the target value in the counterfactual case. After testing this approach on the HELOC (Home Equity Line of Credit) credit application dataset, it was found that the counterfactual explanations generated with feature importance are on average 11% smaller than the explanations without this value.

Goyal et al. (2019) focus on visual counterfactual explanations for image classification. Given an image I which is classified as Y, the counterfactual visual explanation consists of a change to I that will result in the system classifying the changed image as the counterfactual class Y'. This explanation is computed with the following procedure. First a so-called distractor image I' is selected from the database. This distractor image is a random image that is classified as Y' by the system. Secondly, a region in I and in I' is selected by a greedy algorithm in such a way that if the region in I is replaced by the selected region in I', I would be classified as Y'. The explanation is presented to the user by giving both images I and I' where the selected regions are highlighted. For example, for a bird classification problem the selected regions of the original image and the distractor image are the distinctive features of the bird such as the beak.

3.3 Explanation methods for Bayesian networks

As mentioned in paragraph 3.1.3, there are three aspects of a Bayesian network that can be explained; the model itself, the evidence and the reasoning. How the model is explained by visualizing the graph is given in Section 3.3.1. In Section 3.3.2 we show how the most probable and the maximum a posteriori explanation are used to explain the evidence in the model. An overview of different methods to explain the reasoning of a BN is given in Section 3.3.3.

3.3.1 Explaining the model

One of the earliest ways of explaining the model is by giving a visualization of the graph to the user with all associated probabilities. An example of a tool that can do this is HUGIN (Andersen et al., 1989). There are methods that try to give more insight in the graph by giving additional visual clues. Elvira is an example of such a system (Lacave et al., 2006). Elvira colors the arrows in the graph based on the influence of the parent on its child. For example, an arrow from parent to child is colored red if a higher value for the parent makes higher values for the child more likely. Elvira also has a way of computing the magnitude of the influence of an arrow. Arrows are given a thickness that is proportional to this magnitude.

Explanation methods such as HUGIN and Elvira help BN experts to better understand and improve the network. However, these explanations are probable not useful for a user that is unfamiliar with BNs, because he would not

understand the independences and probabilities that are represented in the graph. The explanation method of Druzdzel (1993) provides the user with textual statements in addition to the graph. One assumption he makes in generating these textual explanation is that the network is causal. Examples of textual explanations Druzdzel provides are; *Cold very commonly (p=0.9) causes sneezing* or *There is no other cause of sneezing than cold and allergy.* Conditional dependence can also be expressed to the user, for example with the following sentence; *Given sneezing, cold and allergy are dependent.*

Above examples of explaining the whole model, are easy to understand if the model is small. However, if a BN is more complex with a larger number of variables, the complexity of these methods also increases, which makes them harder to understand. However, the explanation methods are still useful in understanding certain subgraphs of the network.

3.3.2 Explaining the evidence

In the remainder of this section $\mathbf{E}$ denotes a set of evidence and $\mathbf{W}$ denotes a set of unobserved variables.

The aim of explaining the evidence is to give the most probable configuration of $\mathbf{W}$ given $\mathbf{E}$. This configuration of $\mathbf{W}$ is called the most probable explanation (MPE) and the process of obtaining an MPE is called abduction. The MPE can be seen as an explanation of why the evidence has the observed values. However, the explanation does not give a justification as to why the MPE is more probable than other configurations (Lacave & Díez, 2002). The MPE is defined as

$$MPE(\mathbf{e}) = \arg\max_{\mathbf{W}} Pr(\mathbf{w}|\mathbf{e})$$

Notice that an MPE is not necessarily unique, but that there can be multiple configurations that have the same probability. Some abduction methods give just the MPE, other methods are able to give a list of k most probable explanations.

Santos Jr (1991) gives an abduction algorithm that can not only extract the MPE, but also the consecutive most probable configurations of $\mathbf{W}$. The algorithm uses linear constraint systems to compute the MPE. In such a system computations can be done cost efficiently. First, a BN is transformed into a linear constraint system, which can be done in linear time. From the resulting linear constraint system the MPE can be computed. Then the constraint system is updated iteratively and the following most probable configurations are computed from the resulting systems.

A drawback of giving an MPE is that it also assign values to variables that are independent of the evidence. With a maximum a posteriori (MAP) explanation, only a subset of the unobserved variables are assigned a most probable value. With partial abduction an MAP explanation can be generated. Shimony (1991) gives a method of computing a partial abduction on causal networks, where irrelevant variables are not given a value. He gives three different ways of defining independence, which will all cause a variable to be irrelevant. A variable E_i is statistical independent if $P(\mathbf{e}|\mathbf{w}e_i) = P(\mathbf{e}|\mathbf{w})$ for all values e_i of E_i. A variable E_i is δ-independent if $P(\mathbf{e}|\mathbf{w}e_i) - P(\mathbf{e}|\mathbf{w}) \leq \delta$ for all values e_i of E_i. A variable E is quasi-independent if its contribution to the probability of its child is smaller than its own prior probability.

3.3.3 Explaining the reasoning

With the following explanation methods it is assumed there is some target variable T, of which the value needs to be explained. Again $\mathbf{E}$ will denote a set of evidence.

INSITE is one of the first methods to try to explain the reasoning in a network (Suermondt, 1992). Subsequently multiple explanation methods have expanded on this idea. In Section 3.3.3.1 INSITE and methods inspired by INSITE are described. In Section 3.3.3.2 other ideas for explaining the reasoning in the network are given.

3.3.3.1 Inspired by INSITE

INSITE takes two different steps in its explanation. The first step is to identify influential pieces of evidence and to determine if they have a positive or negative influence on T. The second step is to give the chains of reasoning along which the information flows from the influential evidence to the target variable. All methods inspired by INSITE have at least these two steps. We will call giving the outcome of these step to the user the levels of explanation, where the next level expands further on the last.

It is determined if the observation $E_i = e_i$ is influential by looking at the change in probability of the target if E_i was omitted from $\mathbf{E}$. If omitting E_i from $\mathbf{E}$ results in a significant change in the probability of the target, $E_i = e_i$ is considered to be an important observation. Suermondt calls this the cost of omission, which is defined

as: $H(P(T|\mathbf{e}), P(T|\mathbf{e}'))$, where $\mathbf{e}'$ is the observed value configuration for $\mathbf{E}' = \mathbf{E} \backslash \{E_i\}$ and H is the cross entropy. If the cost of omission of E_i is higher than a predefined threshold, then the observation of E_i is significant.

Significant evidence E_i can be consistent or conflicting with $\mathbf{E}$. To determine if E_i is conflicting or consistent the direction of change of the target is used. A direction of change from $P(X)$ to $P'(X)$ states how the probability for every value of X is changed; increased, decreased or not changed. If the direction of change from $P(T|\mathbf{e})$ to $P(T|\mathbf{e}')$ is the same as the direction of change from $P(T|\mathbf{e})$ to $P(T)$ and the cost of omission of E_i is smaller than the cost of omission of $\mathbf{E}$, then E_i is said to be consistent with $\mathbf{E}$. Otherwise, E_i is conflicting. The first level of explanation is to give all significant evidence to the user, with stating for the individual pieces of evidence if they are consistent or conflicting.

The next step in INSITE is to give the chains of reasoning to the user. To determine these chains of reasoning, INSITE first removes all irrelevant nodes from the network. Then all chains from all significant evidence nodes to T are determined. Chains that are insignificant are eliminated. A chain is insignificant if it contains a variable X, that has at most one incoming arc or has no evidence node as successors, and where the cross-entropy between $P(X|\mathbf{e})$ and $P(X)$ is smaller than a predefined threshold. With the help of arc-removals it is determined for all remaining chains if they are conflicting or consistent with $\mathbf{E}$ in a similar way as described above. The second step in the explanation is to present the significant chains of reasoning to the user.

BANTER uses the same underlying ideas as INSITE to provide an explanation (Haddawy et al., 1997). Just like INSITE, BANTER's first step of explanation is to give a list of influential evidence and the second step is to give different chains of reasoning from the evidence to the target. However, how BANTER computes influential evidence and the chains of reasoning differ from how this is done by INSITE.

The influence of evidence on the target is computed with the concept of information. Information is defined in the following way: $I(a_j, b_j) = log(P(a_j|b_j)/P(a_j))$. If the information has a positive value, event $B = b_j$ increases the probability of event $A = a_j$. Otherwise, $B = b_j$ decreases the probability of event $A = a_j$.

If observing $E_i = e_i$ has the same effect on target value t_j as observing the values for $\mathbf{E}$, multiplying the values $I(t_j, e_i)$ and $I(t_j, \mathbf{e})$ will result in a positive value. By summing over all possible values t_j for T, it is computed if the overall effect of observing the value for E_i on the target is the same as the overall effect of observing the values for $\mathbf{E}$. The value resulting from this summation is called the influence of E_i and is computed by $\sum_{t_j \in T} I(t_j, \mathbf{e}) \cdot I(t_j, e_i)$. BANTER also makes a distinction between consistent and conflicting evidence, however Haddawy et al. call this agreeing or disagreeing evidence. If the influence of E_i has a positive value, then E_i agrees with the overall change in probability of the target caused by $\mathbf{E}$. Otherwise, E_i disagrees with $\mathbf{E}$. E_i is said to strongly agree with $\mathbf{E}$ if dividing the influence of this evidence by the largest influence value is higher than a threshold given by the user. Computing if E_i is strongly disagreeing is done in a analogous way.

The first level of explanation now consists of listing all evidence that is strongly agreeing or strongly disagreeing with the whole set of evidence. Note that BANTER uses an opposite idea of INSITE to compute the influence of evidence E_i. Where INSITE looks at the difference in probability of the target given the observed values for $\mathbf{E}$ and the observed values for $\mathbf{E} \setminus \{E_i\}$, BANTER considers the difference between observing values for all evidence in $\mathbf{E}$ and only observing the value for E_i. This difference makes that the evidence that INSITE gives, is a necessary explanation and BANTER gives evidence that is a sufficient explanation (Haddawy et al., 1997).

The second level of explanation consists of giving the most important chains in the graph along which the information from the strongly agreeing or disagreeing evidence flows to the target node. First all active chains from all evidence nodes, selected in the previous level, to the target node are computed by doing a depth-first search through the network. Because chains that are too long do not give an adequate explanation and to speed up the procedure, the length of the found chains is limited to a value specified by the user, with a default of seven. After determining all chains, the chains are ordered on their strength. For every node N on a chain from evidence E_i to the target, the so called impact value is computed in the following way; $\sum_{n_j \in N} |I(n_j, e_i)|$. The strength of a path is equal to the smallest impact value of a node on the chain. Where INSITE gives all chains which are significant, BANTER gives at most five strongest chains leading from the strongly agreeing and disagreeing evidence to the target in the second level of explanation. BANTER does not provide all chains, because this will be too much information for the user to process.

An explanation method by Kyrimi et al. (2019) also gives the same first steps of explanation as INSITE. However, Kyrimi et al. also give a third level of explanation. Just like INSITE, Kyrimi et al. use the cost of omission as a measure of influence of evidence. However, instead of the cross-entropy, the Hellinger distance is used. The Hellinger distance measures the difference between two probability distributions and is symmetric, non-negative, satisfies the triangle inequality and can be computed for both discrete and continuous distributions (Kyrimi et al., 2019).

A threshold is used to determine what evidence variables are significant. Where INSITE uses a static threshold

determined by the user, in this study the threshold is recursively determined in such a way that at least half of all evidence is considered significant. All evidence with an impact above the computed threshold is said to be significant.

Determining if an evidence variable is consistent or conflicting with the whole set of evidence is done by computing $\Delta_t(e_i)$ for every state t of the target variable, with $\Delta_t(e_i) = P(t|\mathbf{e}) - P(t|\mathbf{e}')$ given $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \backslash \{E_i\}$. Note that Δ_t is not defined as an absolute difference, but has a positive or negative sign. E_i is called consistent if for all target states t the difference $\Delta_t(e_i)$ has the same sign as $\Delta_t(\mathbf{e}) = P(t|\mathbf{e}) - P(t)$. Evidence E_i is called conflicting if for all target states t the difference has the opposite sign as $\Delta_t(\mathbf{e})$. In the other cases E_i is said to be mixed.

The first level of explanation of the method given by Kyrimi et al. (2019) is to list all consistent and conflicting evidence in order of their impact value. The second level of explanation tries to give the flow of reasoning. However, Kyrimi et al. do not give a whole chain as INSITE does, but only gives one step in the flow. To give this step the Markov blanket of a variable is computed. The variables in the Markov blanket of the target node that are not observed and not independent of the target are called the intermediate nodes. Level two of the explanation is to list these intermediate nodes and to give the probability of their most likely value.

The last level of explanation is to state the consistent and conflicting evidence for the intermediate variables. This is done analogously to the first level, where the intermediate variables are now seen as the target variable.

3.3.3.2 Other methods

Just like INSITE, a method developed by Van Leersum (2015) also uses different levels of explanation. However, the idea and the computations behind these levels are different. The first level consists of simply stating the value of the target node with the highest probability. For the second level, a set of intermediate nodes is determined. The intermediate nodes are defined as those variables on which the evidence nodes have a high influence and which have a high influence on the probability of the target. To determine the set of intermediate nodes the Edmonds-Karp algorithm is used. The second level consists of these nodes with the most probable value and its probability. For all intermediate nodes, different clusters of evidence nodes that are most influential to the value of this node are generated. These clusters are ordered based on the concept of mutual information. Level three of the explanation lists for every intermediate node the generated clusters in order of importance.

The Explaining BN Inferences (EBI) method proposed by Yap et al. (2008) exploits the idea that a variable is independent of all other variables in the BN given its Markov blanket. The first step of EBI is to determine the Markov blanket of the target node and transform the network in such a way that all variables in the Markov blanket become a parent of the target node and the joint probability distribution is maintained. Based on the conditional probability table of the target and the context-specific independences among the variables a decision tree is generated. A walk down this decision tree can be performed based on the value of the observed variables in the Markov blanket. The explanation given by EBI consists of listing the variables along this path. If a value of a variable in the Markov blanket is unobserved, the most probable value of this variable is explained by recursively applying EBI with this variable as target.

Shih et al. (2018) give two different classes of explanations that use a subset of the evidence to explain the target variable. These classes of explanation can only be used for a certain type of Bayesian Network classifiers, namely the naive Bayes classifier. For efficient computations, the naive Bayes classifier is transformed into an Ordered Binary Decision Diagram (OBDD). It is assumed that all variables and the decision represented by the OBDD can be either positive or negative. An OBDD is a directed, acyclic graph with one root node and two leaves, which represents a positive or negative value for the target variable. All other nodes in the graph represent a variable of the classifier and have two outgoing edges; one with a positive label and one with a negative label. A walk through the graph can be done by starting at the root node and following the positive edge if the root variable is positive and the negative edge other wise. After repeating this step at every node, a leaf node is eventually reached. If this node is positive, the decision made by the OBDD is positive. Otherwise, the decision is negative. Operations such as disjoining, enumerating or complementing can be done efficiently on OBDDs (Shih et al., 2018).

The first class of explanations given by Shih et al. (2018) is the minimum-cardinality explanation. This explanation gives a minimal subset of $\mathbf{E}$ that is a sufficient explanation for the current decision. After transforming a naive Bayes classifier into a OBDD, a minimum-cardinality explanation can be computed with a simple algorithm.

The second class of explanations is that of the prime-implicant explanation. A prime-implicant explanation is a subset of evidence variable that renders the values of all other evidence irrelevant (Shih et al., 2018). So all other evidence variables could be set to an arbitrary value, without changing the decision. This explanation is computed by recursively setting the values of evidence variables to a different value and checking what value for the target is the most probable.

Timmer (2017) has developed an explanation method specifically for a legal domain, however it can also be used for other domains. He tries to extract arguments from the network. This is done with the following procedure. First a so-called support graph is generated from the network. A support graph is a directed acyclic graph that captures the simple chains in the BN that end with the target variable. The idea behind this graph is that all evidence that is entered in the BN, will flow along one of the paths in the support graph. The support graph is then pruned based on the available evidence, after which arguments are generated from the resulting graph.

3.4 Contrasting general AI and BN explanation methods

A common way for explaining BNs and other AI methods is by giving the variables that were most influential for the result of the model. For general AI methods this can for example be done with the Shapley values, for BNs this can be done with the INSITE method. The BN approaches are often more insightful for the following reasons. First of all, they give evidence that is conflicting with the whole set of evidence, where general AI methods only give the consistent input variables of the system. Secondly, the BN methods also provide the user with the chains of reasoning, where general AI system do not provide the reasoning processes. The reasoning processes in for example a neural network are harder to explain, because the reasoning flows through nodes in hidden layers. These nodes do not represent a real world variable and can hardly be explained to a non AI expert. So giving insight in these reasoning processes is a very difficult task and is skipped by most explanation methods.

Rule generation is used as a global explanation for both BNs and other AI systems. An example of this are the textual explanations of Druzdzel for the BN systems and the rule extraction methods given in Section 3.2.2 for AI systems. A difference between the two is that the rule extraction methods for general AI systems can also be used to generate output, where this is not the case for explanations of Druzdzel.

A way of explaining AI systems is with the use of interpretable surrogate models for a local decision of the whole model. This idea is not used with BNs. However, there are some explanation methods such as EBI and the explanation methods given by Shih et al. that transform the graph of the BN, before an explanation is generated. With these methods the BN is transformed into a decision tree and an OBDD respectively. Both these models are interpretable and can thus be used as a surrogate model along with the generated explanation.

The counterfactual explanation is a relatively new idea in the explanation methods for AI systems and has only been researched in the last few years. After our analysis, we could not find any mention of counterfactual explanations for Bayesian networks.

Miller states that people intuitively ask for contrastive explanations. However the explanation methods for BNs and for general AI systems we explored in our literature research, did not contrast one value for the target variable of a system against another value.

4 Contrastive, counterfactual explanations

From the literature research described in Section 3 we concluded that contrastive as well as counterfactual explanations are unexplored types of explanations for BNs. In this section we try to give a definition for contrastive, counterfactual explanations to bridge this gap in the existing literature. It also follows from the research that people generally have a preference of local over global explanations. So we want our definition of a contrastive, counterfactual explanation to explain how certain observations relate to the target variable. As a result we assume in the remainder of this section that there is always a set of observed evidence $\mathbf{E}$, based on which the most probable value for a target T is computed.

To give a definition of a contrastive, counterfactual explanation, we first need to understand what an explanation entails that is only contrastive or only counterfactual. We first define contrastive and counterfactual explanations separately based on different sources in literature. Afterward we combine these definitions.

According to Wachter et al. (2017) a counterfactual is a statement of how an observation would have to be different to result in a desired target value. They state that the counterfactual can be used as an explanation to the user of a system and can tell the subject how to change his behaviour for the system to result in a desired target value. Based on this observation, we give a definition of a counterfactual explanation. However, instead of desired target value, we speak of an expected target value.

Temperature		< 25		25 - 30		> 30	
Sunny		True	False	True	False	True	False
Friend	True	1	0	1	1	1	1
	False	0	0	1	0	1	1

Table 1: This table tells us if we go to the pool based on the variables *Temperature*, *Sunny* and *Friends*. A 1 denotes we go to the pool and a 0 denotes we stay at home.

Definition 4.1. (Counterfactual explanation) Given a set of evidence $\mathbf{E}$ and target T with $\top(T|\mathbf{e}) = t$ and expected value t' with $t \neq t'$, a counterfactual explanation $\mathbf{c}$ is a configuration for a subset $\mathbf{C} \subseteq \mathbf{E}$ where $\mathbf{c}$ has the following properties:

- For each variable in $\mathbf{C}$, its value in $\mathbf{c}$ is other than observed.
- $\top(T|\mathbf{ce}') = t'$, with $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.
- There is no configuration $\mathbf{c}'$ for $\mathbf{C}' \subset \mathbf{C}$ and $\mathbf{c}'$ consistent with $\mathbf{c}$ that meets the previous two conditions.

Given a counterfactual explanation $\mathbf{c}$, we call $\mathbf{C}$ the counterfactual set. Note that if all evidence in $\mathbf{E}$ is binary valued, each counterfactual set has only one corresponding counterfactual explanation, because all evidence variables only have one unobserved value. It now follows that no subset $\mathbf{C}' \subset \mathbf{C}$ is a counterfactual set if there is a counterfactual explanation $\mathbf{c}$ for $\mathbf{C}$. If the evidence in $\mathbf{E}$ is not binary valued, one counterfactual set can have multiple corresponding counterfactual explanations. It also follows that $\mathbf{C}$ and a subset $\mathbf{C}' \subset \mathbf{C}$ can both be counterfactual sets, as long as counterfactual explanations $\mathbf{c}$ and $\mathbf{c}'$ for $\mathbf{C}$ and $\mathbf{C}'$ are not consistent.

A counterfactual explanation can be given to the user in a textual way by stating the variables in $\mathbf{C}$ with the values they are assigned in $\mathbf{c}$ and the expected target value t'.

According to Miller (2019), a contrastive explanation gives reasons why a target value t was more probable than another target value t'. Based on this statement we conclude that a contrastive explanation should explain how the evidence relates to the most probable target value and how it relates to the expected target value. We formalize this in the following way:

Definition 4.2. (Contrastive explanation) Given a set of evidence $\mathbf{E}$ and target T with $\top(T|\mathbf{e}) = t$ and expected value t' with $t \neq t'$, a contrastive explanation contains the following:

- An explanation of how $\mathbf{E}$ relates to target value t.
- An explanation of how $\mathbf{E}$ relates to value t'.

The first requirement can be met in multiple ways. One way is by giving the user a subset of evidence that make all other observations irrelevant. We call the observed value configuration for such a subset a sufficient explanation. We define this in the following way;

Definition 4.3. (Sufficient explanation) Given a set of evidence $\mathbf{E}$ and target T with $\top(T|\mathbf{e}) = t$ and expected value t' with $t \neq t'$, a sufficient explanation is the observed value configuration $\mathbf{s}$ for subset $\mathbf{S} \subseteq \mathbf{E}$, with the following properties;

- $\top(T|\mathbf{se}') = t$ for all possible value configurations $\mathbf{e}'$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$.
- $\mathbf{S}$ is a minimal set for which the first statement holds.

Given a sufficient explanation $\mathbf{s}$, we call the set $\mathbf{S}$ the sufficient set. The sufficient explanation explains how the evidence relates to the target by giving the user that combination of evidence variables that was most decisive for the most probable value for the target. It follows that a sufficient explanation meets the first requirement for a contrastive explanation. However, it does not meet the second requirement, because the sufficient explanation does not explain how the evidence relates to t'.

Just like the definition for a contrastive explanation, the definition for a counterfactual explanation contains a contrast between the most probable value t and an expected value t' for a target T. A counterfactual explanation meets the second requirement, because it relates the evidence to the target by giving those evidence variables that

need to be changed to make the t' the most likely value for T. However, a counterfactual explanation does not meet the first requirement, because it does not give any information about the t.

We explore if combining a counterfactual explanation with a sufficient explanation gives an explanation that is both contrastive and counterfactual. We explore this with the swimming pool example from Section 2.2. With this example we use different observations to determine how likely it is that we enjoy our day at the pool. Based on this probability we decide whether to go to the pool or not. The decision to go to the pool based on all possible observations is given in Table 1.

Example 4.1. Consider the following case in the swimming pool example: One day it is 20 degrees, our friend is going to the pool and it is not sunny. Based on these observations, we decide not go to the pool. We notice that if only the values $Temperature < 25,\ Sunny = false$ were known, *Friend* can take on an arbitrary value without changing the most probable value for *Pool*. It follows that $Temperature < 25 \wedge Sunny = false$ is a sufficient explanation for our decision if no subset of $\{Temperature, Sunny\}$ is a sufficient set. The observation $Sunny = false$ in itself is not a sufficient explanation, because only knowing $Sunny = false$ gives different values for *Pool* as most probable based on the observed values of the other variables. The same reasoning applies to the observation $Temperature < 25$. We conclude that no subsets of $\{Temperature, Sunny\}$ are sufficient sets, so $Temperature < 25,\ Sunny = false$ is a sufficient explanation.

If somebody asks why we did not go to the pool today, we can give two counterfactual explanations. The first explanation we give is $Sunny = true$. This is a correct counterfactual explanation which explains what variables need to be changed to obtain a different result. However, an explanation containing only the variable *Sunny* is not contrastive, because the observed value of *Sunny* in itself is not a sufficient explanation.

The second counterfactual explanation in this case is $Temp > 25$. Again, only providing the user with the variable *Temperature* is not contrastive, because the observed value of *Temperature* is not a sufficient explanation for $Pool = false$.

We now want to give an explanation that is both contrastive and counterfactual. As stated before the observation $Temperature < 25 \wedge Sunny = false$, is a sufficient explanation why we did not go to the pool. Combining this explanation with the counterfactual explanations, gives us the two contrastive, counterfactual explanations; The first explanation is expressed as; *We did not go to the pool today, because the temperature was below 25 degrees and it was not sunny. If the temperature had been higher than 25 degrees, we would have gone to the pool.* We conclude that this explanation is both contrastive and counterfactual, because the first statement tells what evidence was most important for the most probable value of the target and the second statements tells what variables needs to be changed to make the expected value the most probable value of the target.

The second contrastive, counterfactual explanation is; *We did not go to the pool today, because the temperature was below 25 degrees and it was not sunny. If it had been sunny, we would have gone to the pool.* This explanation is both contrastive and counterfactual according to the same reasoning as given above.

Based on this last conclusion we can give a definition of a contrastive, counterfactual explanation.

Definition 4.4. (Contrastive, counterfactual explanation) Given a set of evidence $\mathbf{E}$ and target T with $\top(T|\mathbf{e}) = t$ and expected value t' with $t \neq t'$, a contrastive, counterfactual explanation consists of configurations $\mathbf{s}$ and $\mathbf{c}$ for subsets $\mathbf{S} \subseteq \mathbf{E}$ and $\mathbf{C} \subseteq \mathbf{E}$ respectively such that;

- $\mathbf{s}$ is a sufficient explanation.
- $\mathbf{c}$ is a counterfactual explanation.

Given evidence that is not binary valued, we note that the complexity of finding the sufficient and counterfactual set lies in different aspects of the definitions. The complexity in finding a sufficient explanation lies in verifying if a set is sufficient. Given a evidence $\mathbf{E}$ and a set $\mathbf{S} \subseteq \mathbf{E}$, to verify if $\mathbf{S}$ is potentially a sufficient set a maximum of k^{n-m} different modes need to be computed, with k the maximum number of values a variable in $\mathbf{E}$ can take on, n the size of $\mathbf{E}$ and m the size of $\mathbf{S}$. If it turns out that $\mathbf{S}$ is potentially a sufficient set, it also needs to be verified that $\mathbf{S}$ is minimal.

The complexity in finding all counterfactual explanation lies in computing for all unobserved value configurations if they are counterfactual explanations. Given a set evidence $\mathbf{E}$ and a set $\mathbf{C} \subseteq \mathbf{E}$, to find all unobserved value configurations for $\mathbf{C}$ that are potentially counterfactual explanations $(k-1)^m$ different modes needs to be computed, with k the maximum number of values a variable in $\mathbf{C}$ can take on and m the size of $\mathbf{C}$. Afterwards it also needs to be verified for all found value configurations if they are not consistent with a value configuration for a subset of variables that is a counterfactual explanation.

5 A naive approach of computing the explanation

After we gave a definition for a contrastive, counterfactual explanation we want to construct an algorithm that finds all explanations given a BN with target T and evidence $\mathbf{E}$. Before inference algorithms are performed on a BN, the network is usually pruned by removing all nodes that are irrelevant for the computation from the graph. Baker & Boult (2013) describe a method of how a BN can be pruned. In the remainder of the thesis we will always assume that a BN is pruned before we run the algorithms for finding all sufficient and counterfactual sets. We assume that $\mathbf{E}$ gives the evidence variables that are present in the BN after pruning.

In this section we start with a naive approach of finding all contrastive, counterfactual explanations. We do this for a simple version of the problem, where all evidence is binary valued. Given a BN with a set of evidence $\mathbf{E}$, all sufficient and counterfactual sets are subsets of $\mathbf{E}$. As a result, the most obvious way of computing the explanation, is by checking for each subset of $\mathbf{E}$ if it meets the requirements for a sufficient or counterfactual set. We give algorithms for how all sufficient and counterfactual sets are computed for binary valued evidence in Sections 5.1 and 5.2 respectively. These algorithms give valid explanations, but not necessarily in a time or cost efficient way. An example of how these algorithms work in practice is given in Section 5.3. Afterward we evaluate which parts can be done in a more efficient manner in Section 5.4.

5.1 Computing the sufficient set

Given evidence $\mathbf{E}$ with target T with most probable value t, a set $\mathbf{S}$ needs to meet the following two requirements to be a sufficient set; $\top(t|\mathbf{se}') = t$ for all value configurations $\mathbf{e}'$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ and $\mathbf{S}$ is a minimal set that meets this first requirement. So if we compute all sufficient sets by looping through all subsets of $\mathbf{E}$ we need to check those two requirements.

Algorithm 1 demonstrates how all sufficient sets are found given $\mathbf{E}$ with only binary valued evidence. The algorithm loops over all subsets in ascending order of the size of the subsets. For each subset $\mathbf{S}$ we first check if $\mathbf{S}$ is a superset of a sufficient set that is already found. If this is the case, we know $\mathbf{S}$ is not a sufficient set, so

Algorithm 1: Computes all sufficient sets given a BN with binary valued evidence

```
Input  : A BN with evidence E that is binary valued and target T with most probable value t
Output: All sufficient sets
SufficientSets = ∅
AllSubsets = all subsets S ⊆ E in ascending order of size
foreach S ∈ AllSubsets do
    IsSufficient = true
    foreach S' ∈ SufficientSets do
        if S' ∩ S == S' then
            IsSufficient = false
            break
        end
    end
    if IsSufficient then
        E' = E \ S;
        foreach value assignment e' of E' do
            if P(T|se') ≠ t then
                IsSufficient = false
                break
            end
        end
        if IsSufficient then
            SufficientSets = SufficientSets ∪ S
        end
    end
end
return SufficientSets
```

we continue to the next subset without computing the probability of t given the observed values for $\mathbf{S}$ and the unobserved values for $\mathbf{E} \setminus \mathbf{S}$. If $\mathbf{S}$ is not a superset of a sufficient set, a mode is computed with a standard algorithm for computing probabilities in a BN. If we find that $\mathbf{S}$ is indeed sufficient, we store this set and we continue looking for more subsets.

Algorithm 1 returns all sufficient sets. The sufficient explanations corresponding to these sets are the observed value configurations for those sets.

5.2 Computing the counterfactual set

Computing all counterfactual sets is done in a way similar to computing all sufficient sets. Because we only consider binary valued evidence, all evidence only has one unobserved value. It follows that one counterfactual set only gives one counterfactual explanation and no subset of a counterfactual set is a counterfactual set as well. A similar way as used to find all sufficient sets can be used to find all counterfactual sets. This is described in Algorithm 2. This algorithm returns all counterfactual sets $\mathbf{C}$, the corresponding counterfactual explanations are the value configuration where all evidence variables in $\mathbf{C}$ take on their unobserved value.

Algorithm 2: Compute all counterfactual sets given a BN with evidence

Input : A BN with evidence $\mathbf{E}$ that is binary valued and target T with expected value t'
Output: All counterfactual sets

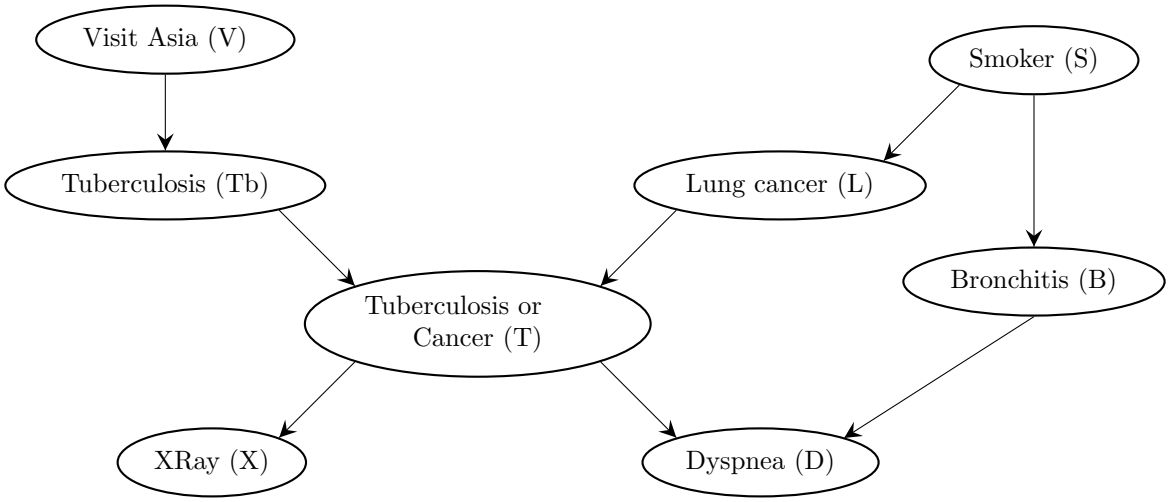
```
1  CounterfactualSets = ∅
2  AllSubsets = all subsets S ⊆ E in ascending order of size
3  foreach C ∈ AllSubsets do
4  |  IsCounterfactual = true
5  |  foreach C' ∈ CounterfactualSets do
6  |  |  if C' ∩ C == C' then
7  |  |  |  IsCounterfactual = false
8  |  |  |  break
9  |  |  end
10 |  end
11 |  if IsCounterfactual then
12 |  |  e'= unobserved value configuration for E' = E \ C
13 |  |  if ⊤(T|ce') = t' then
14 |  |  |  CounterfactualSets = CounterfactualSets ∪ C
15 |  |  end
16 |  end
17 end
18 return CounterfactualSets
```

5.3 Example

We will further explain the algorithms above with the Visit Asia network given by Lauritzen & Spiegelhalter (1988). This network is given in Figure 2. This BN can be used to determine how probable it is that a subject has tuberculosis or lung cancer. So we consider *Tuberculosis or Cancer* (TC) as the target. The variables *VisitAsia* (V), *XRay* (X), *Smoker* (S) and *Dyspnea* (D) are observable. All these variables are binary valued and can take on the values $true$ or $false$. The remaining variables are hidden, so these do not take part in the computations.

Example 5.1. Consider the case with the following evidence; The subject has visited Asia, he is a smoker, his XRay results are abnormal and he has dyspnea. We have the expectation that the patient does not have tuberculosis or lung cancer, but given this evidence it is more probable that the patient has tuberculosis or lung cancer; $P(tc|vxsd) = 0.8398$. We first compute the sufficient sets with Algorithm 1.

The algorithms loops through all subsets in ascending order of size. We start with the empty set. Because no sufficient sets are yet found we skip the loop starting at line 5. The most probable value for all value configurations of $\{VisitAsia, XRay, Smoker, Dyspnea\}$ needs to be checked. Not all those configurations have $TC = true$ as the most probable value. As soon as the first configuration for which this is the case is checked, we break out of the loop. The current set is not sufficient, so we do not add it to all sufficient sets.



$P(v) = 0.01$ $P(s) = 0.5$

$P(tb\|v) = 0.05$	$P(b\|s) = 0.6$	$P(l\|s) = 0.1$
$P(tb\|\neg v) = 0.01$	$P(b\|\neg s) = 0.3$	$P(l\|\neg s) = 0.01$
$P(x\|t) = 0.98$	$P(d\|b,t) = 0.9$	$P(t\|l,tb) = 1$
$P(x\|\neg t) = 0.05$	$P(d\|\neg b,t) = 0.7$	$P(t\|\neg l,tb) = 1$
	$P(d\|b,\neg t) = 0.8$	$P(t\|l,\neg tb) = 1$
	$P(d\|\neg b,\neg t) = 0.1$	$P(t\|\neg l,\neg tb) = 0$

Figure 2: The Visit Asia network with evidence variables $VisitAsia, XRay, Smoker$ and $Dyspnea$ and target $TuberculosisorCancer$

The same steps are taken for the next subsets. $\{XRay, Smoker, Dyspnea\}$ is the first sufficient set we find. We add it to the set of all sufficient sets. Because a sufficient set is found, we need to check for the following subsets if it is not a superset of this sufficient set. In this way we find that $\{VisitAsia, XRay, Dyspnea\}$ and $\{VisitAsia, XRay, Smoker\}$ are also sufficient sets. For the last subset, $\{VisitAsia, XRay, Smoker, Dyspnea\}$, we notice in the loop starting at line 5 that this is a superset of all of the sufficient sets already found. We conclude that $\{VisitAsia, XRay, Smoker, Dyspnea\}$ can not be a sufficient set, so we do not enter the loop at line 13. All subsets are checked now and we found all three sufficient sets.

We also want to find all counterfactual sets. We loop over all subsets in the same ways as before. We again start with the empty set. Because no counterfactual sets are found, we skip the loop at line 5. Because there are no value configurations for an empty set the loop at line 13 is skipped. We now move to the next subset $\{Dyspnea\}$. Again we skip the loop at line 5. It is computed that the target does not have the expected value given the unobserved value for $Dyspnea$ and the observed values for $\{VisitAsia, XRay, Smoker\}$, so $\{Dyspnea\}$ is not a counterfactual set. We continue the search and the first subset that is found is $\{XRay\}$. It follows that for all supersets of $\{XRay\}$ no probability needs to be computed to conclude that it is not a counterfactual set. When the algorithm is finished, the following counterfactuals sets are found; $\{XRay\}$, $\{Smoker, Dyspnea\}$, $\{VisitAsia, Dyspnea\}$ and $\{VisitAsia, Smoker\}$. The corresponding counterfactual explanations are the value configurations where all values in the counterfactuals sets take on their unobserved value.

All contrastive, counterfactual explanations are now constructed by combining all sufficient and counterfactual explanations.

5.4 Observations and improvements

As stated previously, the methods described above are valid ways of computing sufficient and counterfactual sets, however they are not an efficient way of doing so. A set of size n has 2^n subsets, so the loop starting at line 3 for both algorithms is started 2^n times for an evidence set of size n.

In Algorithm 1 there is another loop at line 13 over all possible value configurations of a certain subset. There are k^m iterations in this loop in the worst-case scenario, where subset $\mathbf{E}'$ has size $m \leq n$ and the variables in the subset have k possible values. Inside this last loop a probability is computed. Probabilistic inference also has an exponential time complexity. Also note that the mode of the target given one value configuration is computed multiple times in the computations for the sufficient sets. It must be obvious that computing sufficient sets with the method discussed above is infeasible in practical situations. Algorithm 2 has a lower complexity, because there is no loop over all value configurations.

We also note that Algorithms 1 and 2 are performed consecutively. As a result, some probabilities are computed multiple times when finding all sets. We want to improve the current methods for finding all sufficient and counterfactual sets by constructing algorithms that compute a probability for a specific value configuration at most once. We also want the algorithms for finding all sufficient and counterfactual sets that can be easily combined to find all explanations in one run.

6 Using lattices to compute the explanation

In this section we explore how a lattice can be used to more improve Algorithms 1 and 2. Again we assume that the network is pruned before we try to find all explanations. We also only consider situations where all evidence variables are binary valued.

In Section 6.1 it is explained how each subset in a subset lattice can be labeled with the most probable value for the target given a certain value configuration of all evidence. How these labels are used to compute all sufficient and counterfactual sets from the lattice is described in Section 6.2 and 6.3 respectively. This is all clarified with an example in Section 6.4. We conclude with some observations and possible improvements in Section 6.4.1.

6.1 The lattice

A lattice is a partially ordered set where every two finite, nonempty subsets have a unique least upper bound and a unique greatest lower bound (Grätzer, 2011). Given a set of elements $\mathbf{P}$, all subsets of $\mathbf{P}$ can be partially ordered by subset inclusion, which results in the subset lattice. The lattice is an undirected graph structured in layers; a layer i contains all subsets of size i. The lowest layer is layer 0 and only contains the empty set. We also call this the bottom of the lattice. The highest layer is also called the top of the lattice and contains the complete set. All subsets in layer i are connected with an edge to the nodes in layer $i+1$ with which they have a subset relation. Given a node X in layer i, the children of X are the subsets in layer $i-1$ that share an edge with X. The parents of X are the subsets in layer $i+1$ that share an edge with X. The ancestors of X are all nodes located on the paths from X to the top of the lattice, where each consecutive node is situated in a higher layer. The descendants of X are all nodes located on the paths from X to the bottom of the lattice, where each consecutive node is situated in a lower layer.

We can use the subset lattice described above to represent all subsets of the evidence set. We want to use this lattice to more efficiently compute all sufficient and counterfactual sets from it. To be able to do this we enhance the subset lattice by labeling each subset in the lattice with a certain probability. This leads to the following definition of the enhanced subset lattice.

Definition 6.1. (Enhanced subset lattice) Given an evidence set $\mathbf{E}$ with binary valued evidence, a target T with $\top(T|\mathbf{e}) = t$ and an expected value t' with $t' \neq t$, the enhanced subset lattice is the subset lattice of $\mathbf{E}$ where each subset $\mathbf{S} \subseteq \mathbf{E}$ is labeled with $\top(T|\mathbf{se}')$ with $\mathbf{s}$ the value configuration where all evidence variables in $\mathbf{S}$ take on their observed values and $\mathbf{e}'$ the value configuration where the evidence variables in $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ take on their unobserved values.

Note that this definition places the restriction on the evidence that it is all binary valued. The target does not necessarily have to be binary. Given a set $\mathbf{S} \subseteq \mathbf{E}$ in the lattice, we call $\mathbf{C}$ with $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ the inverse of $\mathbf{S}$.

An example lattice can be found in Figure 3. This is a lattice of the Visit Asia network given in Figure 2. The variable *Tuberculosis or Cancer* (T) is the target. *Visit Asia (V), XRay (X), Smoker (S)* and *Dyspnea* (D) is the evidence in the network. We have the observations that the patient has visited Asia, is a smoker, has dyspnea and

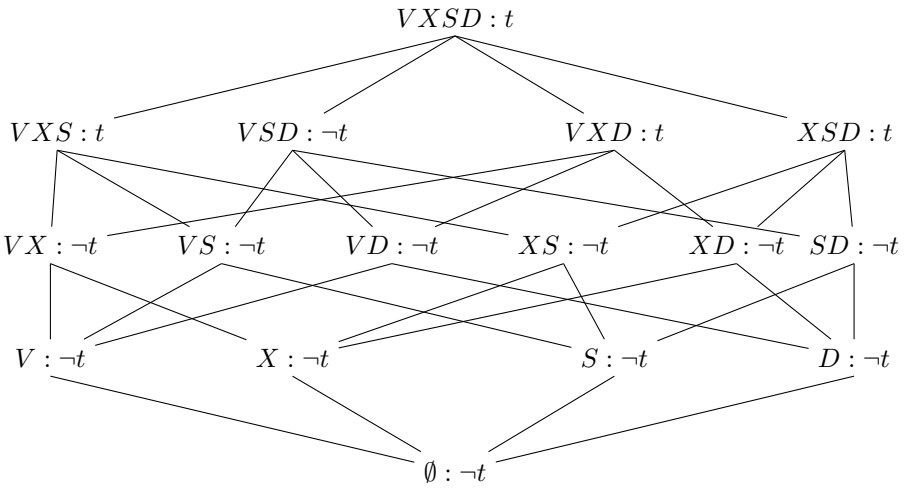


Figure 3: A subset lattice representing the evidence in the Visit Asia network. The labels are computed for the following observations; *VisitAsia = true, Smoker = true, Dyspnea = true, XRay = true*

has an abnormal XRay. The nodes in the lattice have all the appropriate label given this evidence. For example, the node containing the subset $\{V, S, D\}$ is denoted with $\neg t$, because it is most likely that the patient does not have tuberculosis or lung cancer given this evidence; $\top(T|vsd\neg x) = \neg t$.

The enhanced lattice now contains enough information to compute all sufficient and counterfactual sets. We first explain how it can be determined if a set is sufficient, afterward we explain the same thing for the counterfactual sets.

6.2 Deriving sufficient sets from the lattice

Given evidence $\mathbf{E}$ with only binary valued evidence and most probable value t of target T, only the subsets that are labeled with t in the lattice can be a sufficient set. A subset $\mathbf{S}$ that is not labeled with t can not be sufficient, because it follows from Definition 6.1 that t is not the most probable value for T given the observed values for $\mathbf{S}$ and a value configuration for $\mathbf{E} \setminus \mathbf{S}$. Not all sets labeled with t are sufficient. For example, the top of the lattice gives the whole evidence set and is therefore is always labeled with t. However, this set is not necessarily a sufficient set, because it does not have to be minimal.

To decide if a subset $\mathbf{S}$ is sufficient, we first need to know if t is the most probable value for T given the observed values for $\mathbf{S}$ and all possible value configurations $\mathbf{e}'$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. The following proposition states how we can find the most probable values for these value configurations in the lattice;

Proposition 6.1. Given an evidence set $\mathbf{E}$ with binary valued evidence, a target T with $\top(T|\mathbf{e}) = t$ and a subset $\mathbf{S} \subseteq \mathbf{E}$ in the lattice given by Definition 6.1, the values for $\top(T|\mathbf{se}')$ for the observed values of $\mathbf{S}$ and all possible value configurations $\mathbf{e}'$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ are given by the labels of $\mathbf{S}$ and all its ancestors.

Proof. Assume we have a subset $\mathbf{S} \subseteq \mathbf{E}$ in the lattice with $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. It follows from Definition 6.1 that the label for $\mathbf{S}$ itself represents the value configuration where all variables in $\mathbf{S}$ take on their observed value and all variables in $\mathbf{E}'$ take on their unobserved value.

We now need to find the nodes in the lattice that represent the value configurations where the variables in only a subset $\mathbf{E}'' \subset \mathbf{E}'$ take on unobserved values. For all subsets $\mathbf{E}'' \subset \mathbf{E}'$, there is a subset $\mathbf{A} = \mathbf{E} \setminus \mathbf{E}''$. It follows from Definition 6.1, that the label of $\mathbf{A}$ in the lattice gives the most probable value of T given the observed values for $\mathbf{A}$ and the unobserved values for evidence variables in $\mathbf{E}''$. Because $\mathbf{E}'' \subset \mathbf{E}'$, $\mathbf{S} \cup \mathbf{E}' = \mathbf{E}$ and $\mathbf{A} \cup \mathbf{E}'' = \mathbf{E}$, it follows that $\mathbf{S} \subset \mathbf{A}$. It follows from the definition of the subset lattice that $\mathbf{A}$ is an ancestor of $\mathbf{S}$ in the lattice. The label for $\mathbf{A}$ represents the value configuration where all variables in $\mathbf{A}$ take on their observed values and all evidence in $\mathbf{E}''$ take on their unobserved values. We derive that all value configurations for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ are represented by $\mathbf{S}$ and all ancestors of $\mathbf{S}$. So the labels for $\mathbf{S}$ and all of its ancestors give the most probable value of T given the observed values for $\mathbf{S}$ and all value configurations for $\mathbf{E} \setminus \mathbf{S}$. □

From the previous proposition, it follows that t is the most probable value for T given the observed values for $\mathbf{S}$ and all possible value configurations for $\mathbf{E} \setminus \mathbf{S}$ if all these ancestors of $\mathbf{S}$ are labeled with t. We can not yet conclude

that $\mathbf{S}$ is sufficient, because we also need to check if it is a minimal set. The following propositions state how we can derive from the lattice if a set $\mathbf{S}$ is sufficient.

Proposition 6.2. Given an evidence set $\mathbf{E}$ with binary valued evidence, a target T with $\top(T|\mathbf{e}) = t$ and a subset $\mathbf{S} \subseteq \mathbf{E}$ in the lattice given by Definition 6.1, $\mathbf{S}$ is a sufficient set if the following statements hold;

- $\mathbf{S}$ is labeled with t.
- All ancestors of $\mathbf{S}$ are labeled with t.
- All children of $\mathbf{S}$ are not labeled with t or have at least one ancestor that is not labeled with t.

Proof. Assume we have a set $\mathbf{S} \subseteq \mathbf{E}$ in the lattice for which all statements hold. It follows from the first two statements and Proposition 6.1 that $\top(T|\mathbf{se}') = t$ given the observed values for $\mathbf{S}$ and all possible value configurations $\mathbf{e}'$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. If no subset of $\mathbf{S}$ is a sufficient set, it follows that $\mathbf{S}$ is minimal and therefore a sufficient set.

First consider all children $\mathbf{S}'$ of $\mathbf{S}$ in the lattice that are not labeled with t. We drive from the label of $\mathbf{S}'$ and Definition 6.1 that $\mathbf{S}'$ is not a sufficient set. The descendants $\mathbf{D}$ of $\mathbf{S}$ with $\mathbf{D} \subset \mathbf{S}'$ all have an ancestor that is not labeled with t, namely $\mathbf{S}'$. It follows from Proposition 6.1, that given the observed values for $\mathbf{D}$ there is a value configuration for $\mathbf{E} \setminus \mathbf{D}$ where t is not the most probable value for T. So $\mathbf{D}$ is not a sufficient set. It follows that all children of $\mathbf{S}$ that are not labeled with t and all of their descendants are not sufficient sets.

Now consider the children $\mathbf{S}'$ of $\mathbf{S}$ in the lattice that are labeled with t, but have at least one ancestor that is not labeled with t. Because $\mathbf{S}'$ has an ancestor that is not labeled with t, it follows from Proposition 6.1 that for the observed values of $\mathbf{S}'$ there is a value configuration for $\mathbf{E} \setminus \mathbf{S}'$ where t is not the most probable value for T. Therefore $\mathbf{S}'$ is not a sufficient set. All subsets of $\mathbf{S}'$ are also not sufficient sets, because they all share the ancestor of $\mathbf{S}'$ that is not labeled with t.

We now proved for all descendants of $\mathbf{S}$ that they can not be sufficient sets. It follows that $\mathbf{S}$ is minimal and is therefore a sufficient set. □

In conclusion, a subset $\mathbf{S}$ is sufficient if $\mathbf{S}$ and all its ancestors are labeled with t and all children of $\mathbf{S}$ are not labeled with t or have at least one ancestor that is not labeled with t. We find all these sufficient sets with a breadth-first search starting at the top of the lattice. Algorithm 3 gives how this is done while dynamically constructing the lattice. Because the lattice is dynamically constructed, only the necessary probabilities are computed. At the end of the algorithm all sets are returned for which all children are not labeled with t. These children either have t' as label or do not have any label, because it was established the set was not sufficient before this label was computed. Because the children of sets not labeled with t are not explored by the algorithm, the sufficient sets are exactly those sets of which all children are leafs.

The sufficient explanations follow directly from the sufficient sets. Given a sufficient set $\mathbf{S}$, the corresponding sufficient explanation is the observed value configuration for $\mathbf{S}$.

In Section 6.4 an example is given of how this algorithm can be combined with the algorithm for computing the counterfactual sets.

Algorithm 3: Computes all sufficient sets with a breadth-first search through the subset lattice

Input : Target T with most probable value t,
evidence set $\mathbf{E}$ with only binary valued evidence
Output: All sufficient sets

```
Q = Queue containing E
while Q not empty do
    S = get first item in Q
    if S has no parent not labeled with t then
        t' = ⊤(T|se') with observed values in S and counterfactual values in E' = E \ S
        if t' == t then
            Put all children of S in Q
        end
    end
end
return All sets labeled with t for which all children are not labeled with t
```

6.3 Deriving counterfactual sets from the lattice

Where we only consider nodes labeled with t to be potential sufficient sets, we only consider the inverses of nodes labeled with t' to be the counterfactual sets, with $t' \neq t$ the expected value of the target. Given a node in the lattice with subset $\mathbf{S}$ that is labeled with t', its inverse $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ is a potential counterfactual set. The following proposition states when this set is considered to be an actual counterfactual set.

Proposition 6.3. Given an evidence set $\mathbf{E}$ with binary valued evidence, a target T with $\top(T|\mathbf{e}) = t$ and an expected value t' with $t' \neq t$ and a subset $\mathbf{S} \subseteq \mathbf{E}$ in the lattice given by Definition 6.1, the set $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ is a counterfactual set if $\mathbf{S}$ is labeled with t' and none of its ancestors is labeled with t'.

Proof. Assume we have a subset $\mathbf{S} \subseteq \mathbf{E}$ in the lattice that is labeled with t', none of its ancestors is labeled with t' and has inverse $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$. It follows from Definition 6.1 that $\top(T|\mathbf{sc}) = t'$ with $\mathbf{s}$ the observed value configuration for $\mathbf{S}$ and $\mathbf{c}$ the value configuration for $\mathbf{C}$ where all variables take on their unobserved values. Because all evidence in $\mathbf{E}$ is binary valued it follows that $\mathbf{C}$ is a counterfactual set if no subset of $\mathbf{C}$ is a counterfactual set. For a subset $\mathbf{C}' \subset \mathbf{C}$, we define the set $\mathbf{A} = \mathbf{E} \setminus \mathbf{C}'$. Because $\mathbf{C} \cup \mathbf{S} = \mathbf{E}$, $\mathbf{C}' \cup \mathbf{A} = \mathbf{E}$ and $\mathbf{C}' \subset \mathbf{C}$, it follows that $\mathbf{S} \subset \mathbf{A}$. So $\mathbf{A}$ is a ancestor of $\mathbf{S}'$ in the lattice. We derive that the inverses of all ancestor of $\mathbf{S}$ give all subsets of $\mathbf{C}$.

If an ancestors $\mathbf{A}$ is labeled with t', t' is the most probable value for T given the observed values for $\mathbf{A}$ and the unobserved values for $\mathbf{C}'$. In this case $\mathbf{C}'$ could be a counterfactual set. If $\mathbf{A}$ has another label it follows that $\mathbf{C}'$ can not be a counterfactual set.

Because we assumed that all ancestors of $\mathbf{S}$ are not labeled with t', it follows that no subsets of $\mathbf{C}$ are counterfactual sets. We conclude that $\mathbf{C}$ is a counterfactual set. □

All counterfactual sets can be found by a breadth-first search starting at the top of the lattice. During the search, the lattice is dynamically built. In this way only the necessary probabilities are computed. Pseudocode for how all counterfactual sets can be found with a breadth-first search through the lattice is given by Algorithm 4. Because the evidence is all binary valued, the counterfactual explanations follow directly from the counterfactual sets. Given a counterfactual set $\mathbf{C}$, the corresponding counterfactual explanation is the value configuration where all evidence variables in $\mathbf{C}$ take on an unobserved value.

Algorithm 4: Computes all counterfactual sets with a breadth-first search through the subset lattice

```
   Input  : Target T with expected value t' and evidence set E with only binary valued evidence
   Output: All counterfactual sets
 1 CounterfactualSets = ∅
 2 Q = Queue containing E
 3 while Q not empty do
 4 |   S = get first item in Q
 5 |   C = E \ S
 6 |   if S has no parent labeled with t' then
 7 |   |   t = most probable value of P(T|sc)
 8 |   |   if t == t' then
 9 |   |   |   CounterfactualSets = CounterfactualSets ∪ C
10 |   |   else
11 |   |   |   Put all children of S in Q
12 |   |   end
13 |   end
14 end
15 return CounterfactualSets
```

6.4 Example

In this section we demonstrate how Algorithm 3 and 4 can be combined to compute all sufficient and counterfactual sets with one breadth-first search. We use the Visit Asia network for this purpose. We use the example where the patient has visited Asia, is a smoker, has an abnormal XRay and suffers from dyspnea. Given this evidence, the probability that the patient has tuberculosis or lung cancer is 0.83.

The subset lattice can be found in Figure 3. Each subset in the lattice is denoted with the most likely value for the target *Tuberculosis or Lungcancer* (T). These denotations are only given to illustrate the idea behind the lattice. In practice, only the necessary probabilities are computed.

Example 6.1. We start our breadth-first search at the top of the lattice. For breadth-first search we use a queue to keep track of the next subset. We start with a queue containing only the top of lattice with label t. Because t is the most likely value for this subset, we can conclude that its inverse is no counterfactual set. We do not have enough information to know if the set itself is sufficient. So we put the children of the current subset in the queue.

The first subset $\{VisitAsia, XRay, Smoker\}$ is taken from the queue. We find that $\top(T|vxs\neg d) = t$, so $\{VisitAsia, XRay, Smoker\}$ is labeled with t. The parent of the current set is also labeled with t; it follows that the parent can not be sufficient, because it is not minimal. The current subset is potentially a sufficient set. To verify this, we need to check the most probable values for all subsets of $\{VisitAsia, XRay, Smoker\}$. So we put the children of the current subset in the queue.

We continue to the next subset $\{VistAsia, Smoker, Dyspnea\}$. We compute that its label is $\neg t$. It follows that the current subset can not be a sufficient set. Because the only parent of the current subset is labeled with t, it follows from Proposition 6.3 that the inverse of the current set is a counterfactual set. So $\{XRay\}$ is the first counterfactual set we find. We do not put the children of $\{VistAsia, Smoker, Dyspnea\}$ in the queue. The children are not sufficient sets, because they have a parent labeled with $\neg t$. Their inverse can also not be a counterfactual set, because the inverses are supersets of $\{XRay\}$. So there is no need to compute their probabilities.

For the next two subsets in the queue $\{VisitAsia, XRay, Dyspnea\}$ and $\{XRay, Smoker, Dyspnea\}$, we compute the appropriate probabilities and find that in both cases t is the most probable value. We are not yet able to derive if they are sufficient sets, so we put their children in the queue.

We now arrived at the next layer of the lattice with subsets of two elements. Of these subsets $\{VisitAsia, XRay\}$ is the first in the queue. We find that $\neg t$ is its label. It follows that $\{VisitAsia, XRay\}$ is not a sufficient set. All ancestor of $\{VisitAsia, XRay\}$ are labeled with t, so we derive from Proposition 6.3 that the inverse $\{Smoker, Dyspnea\}$ is a counterfactual set. It follows that none of the descendants of $\{VisitAsia, XRay\}$ are sufficient sets and none of the inverses of the descendants are counterfactual sets, so we do not put the children of $\{VisitAsia, XRay\}$ in the queue.

The next subset is $\{VisitAsia, Smoker\}$. We do not need to compute the probability for this subset. One of its ancestors is labeled with $\neg t$, so $\{VisitAsia, Smoker\}$ is not a sufficient set and neither is its inverse a counterfactual set.

For the next subset $\{XRay, Smoker\}$, we again find that $\neg t$ is its label and we can conclude that its inverse $\{VisitAsia, Dyspnea\}$ is a counterfactual set. We now concluded for each child of $\{VisitAsia, XRay, Smoker\}$ that it is not a sufficient set. It now follows from Proposition 6.2 that $\{VisitAsia, XRay, Smoker\}$ is a sufficient set.

The following two subsets in the queue $\{VisitAsia, Dyspnea\}$ and $\{Smoker, Dyspnea\}$ have a parent that is denoted with $\neg t$ so we do not have to compute their probabilities. The last subset in the queue is $\{XRay, Dyspnea\}$. We compute that $\neg t$ is its label. We conclude that its inverse $\{VisitAsia, Smoker\}$ is a counterfactual set. We have now derived that all children of $\{VisitAsia, XRay, Dyspnea\}$ and $\{XRay, Smoker, Dyspnea\}$ are not a sufficient set, either because it does not have t as most probable value or because it has an ancestor that is not labeled with t. It is concluded that both $\{VisitAsia, XRay, Dyspnea\}$ and $\{XRay, Smoker, Dyspnea\}$ are sufficient sets.

The queue is now empty, which means that we computed all necessary probabilities to find all sufficient and counterfactual sets. The sufficient sets we found during the computations are $\{VisitAsia, XRay, Smoker\}$, $\{VisitAsia, XRay, Dyspnea\}$ and $\{XRay, Smoker, Dyspnea\}$. The counterfactual sets we found are $\{XRay\}$, $\{Smoker, Dyspnea\}$, $\{VisitAsia, Dyspnea\}$ and $\{VisitAsia, Smoker\}$. To find these sets a total of 8 probabilities were computed.

Because the evidence was binary valued, the counterfactual explanations follow directly from the counterfactual sets. All sufficient explanations can now be combined with all counterfactual explanations to construct all contrastive, counterfactual explanations.

6.4.1 Observations and improvements

This method of computing all sufficient and counterfactual sets improves the previous methods, because all probabilities are computed at most one time and each probability is used to make conclusions about both the sufficient and counterfactual sets. In the worst-case scenario all 2^n probabilities in the lattice need to be computed, where n is the size of the evidence set. Because the breadth-first search of Algorithm 3 and 4 start at the top of the lattice, only 2^n are computed if the label of the empty set needs to be computed. It follows from Proposition 6.2 and 6.3

that the label for the empty set is only computed if all other sets are labeled with t, with t the most probable value for target T given all evidence. In all other cases, we are able to conclude that the empty set is not sufficient and its inverse is not counterfactual without computing the appropriate probability. After the label for the empty set is computed, we find either that it is sufficient or that its inverse is counterfactual. If the empty set is sufficient, it follows that none of the evidence has any influence on the target. This seems unlikely to often be the case. If the inverse of the empty set is a counterfactual set, the target only has t' as most probable value if all evidence takes on the counterfactual value.

We notice that the number of probabilities to be computed in the lattice is dependent on the size of the sufficient and counterfactual sets. When we start a breadth-first search at the top of the lattice and all sufficient sets have a size close to the size of the whole evidence set, we have to compute relatively few probabilities. If all sufficient sets have a small size, we have to compute relatively more probabilities. It seems logical that starting the search at the bottom of the lattice gives a more efficient algorithm if all sufficient sets have a small size. However, we need to know the label for all ancestors of a subset in the lattice to be able to derive if a set is sufficient. So starting the search at the bottom of the lattice does not reduce the number of probabilities to compute in case of sufficient sets with a small size.

Another drawback of Algorithm 3 and 4 is that they only work for evidence that is binary valued. The algorithms are easily extended to also work for evidence variables that have more than two values. For example, the subsets $\mathbf{S}$ in the lattice can be labeled with *true* if t is the most probable value given the observed values for $\mathbf{S}$ and all unobserved value configurations $\mathbf{e}'$ with $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ and $false$ otherwise. The same breadth-first search can then be performed to find all sufficient and counterfactual sets. However, now k^{n-m} probabilities need to be computed for all subsets $\mathbf{S} \subseteq \mathbf{E}$ in the lattice in comparison to one probability per node for evidence that is binary valued, where k is the maximum number of values a variable in $\mathbf{E}$ can take on, n is the size of $\mathbf{E}$ and m is the size of $\mathbf{S}$. Moreover, finding all counterfactual explanations becomes more complex, because evidence variables can take on multiple unobserved values. It then also follows that a counterfactual set can have a subset that is also a counterfactual set.

We derive that if we had some additional information about the relation of different values of the evidence with the target, we could use this to more efficiently compute all explanations for evidence that is not binary valued. In the next section we explore how monotonicity can be used in this regard.

7 Monotonicity

In this section we explore how a monotonicity relation between the evidence variables and a target can be exploited to more efficiently compute a contrastive, counterfactual explanation. First we introduce two different types of monotonicity in Section 7.1. With these definitions we are able to derive several propositions about the inclusion of an evidence variable in a sufficient or counterfactual set based on the ordering of the observed value for the variable and the ordering of the most probable and expected value for the target. These definitions are given in Section 7.2. Later we demonstrate in Section 8 how these propositions are used to give a new definition for an enhanced subset lattice and to compute all explanations from this lattice.

7.1 Definition of monotonicity

L. Van der Gaag et al. (2004) introduce two kinds of monotonicity for BNs; monotonicity in mode and monotonicity in distribution. The following assumptions are made; for each variable V in the BN we assume that there is a total ordering on the set of possible values V can take on. If V is binary we assume the ordering $\neg v < v$. For each set of variables $\mathbf{V}$ we assume there is a partial ordering $\preceq$ on the joint value assignments of $\mathbf{V}$ given the total ordering on each $V \in \mathbf{V}$.

We first give the definition for monotonicity in mode. The mode output function $f_\top$ is defined in the following way for each possible value configuration $\mathbf{e}$ for $\mathbf{E}$.

$$f_\top = \top(T|\mathbf{e})$$

Monotonicity in mode is now defined in the following way.

Definition 7.1. (Monotonicity in mode)

- The relation between T and $\mathbf{E}$ is called isotone in mode if for all value configurations $\mathbf{e}, \mathbf{e}'$ for $\mathbf{E}$ it holds that

$$\mathbf{e} \preceq \mathbf{e}' \rightarrow f_\top(\mathbf{e}) \leq f_\top(\mathbf{e}')$$

- The relation between T and $\mathbf{E}$ is called antitone in mode if for all value configurations $\mathbf{e}, \mathbf{e}'$ for $\mathbf{E}$ it holds that

$$\mathbf{e} \preceq \mathbf{e}' \rightarrow f_{\top}(\mathbf{e}) \geq f_{\top}(\mathbf{e}')$$

If a piece of evidence is isotone in mode with a target, assigning a higher ordered value can not lead to a lower ordered value that is most likely for the target. The opposite is true for a variable that is antitone in mode. Given a BN with an evidence set $\mathbf{E}$, some evidence in $\mathbf{E}$ could be isotone in mode while other evidence is antitone in mode. If all evidence in a set is either isotone or antitone in mode with a target, we simply say that the whole set is monotone in mode. The isotonicity or antitonicity of a certain variable is further specified when necessary.

Now we define monotonicity in distribution. For all values v of variable V the *cumulative distribution function* F_P is defined as

$$F_P(v) = P(V \leq v)$$

Given two probability distributions $P(V)$ and $P'(V)$, $P'(V)$ is called *stochastically dominant* over $P(V)$ if $F_{P'}(v) \leq F_P(v)$ for all values v of V. Monotonicity in distribution is now defined in the following way.

Definition 7.2. (Monotonicity in distribution)

- The relation between T and $\mathbf{E}$ is called isotone in distribution if for all value configurations $\mathbf{e}$, $\mathbf{e}'$ for $\mathbf{E}$ with $\mathbf{e} \preceq \mathbf{e}'$ it holds that $P(T|\mathbf{e}')$ is stochastically dominant over $P(T|\mathbf{e})$.
- The relation between T and $\mathbf{E}$ is called antitone in distribution if for all value configurations $\mathbf{e}$, $\mathbf{e}'$ for $\mathbf{E}$ with $\mathbf{e} \preceq \mathbf{e}'$ it holds that $P(T|\mathbf{e})$ is stochastically dominant over $P(T|\mathbf{e}')$.

If a variable is isotone in distribution with a target, assigning a higher ordered value for a variable can not lead to a decrease in probability for highest ordered value for the target. The opposite is true for a variable that is antitone in distribution. Again, we say that a set of variables in monotone in distribution if all variables in the set are either isotone or antitone in distribution with a target.

The two concepts of monotonicity represent different properties of a BN. Monotonicity in mode is more useful if only the most likely value of the target is needed, while monotonicity in distribution is most useful if the distribution of the target is needed for further computations. Because they represent different properties, we should consider what properties are the most useful for the computation of a contrastive, counterfactual explanation.

7.2 Propositions derived from monotonicity

In the following sections several propositions are proven about the inclusion of evidence variables in the sufficient or counterfactual sets given a monotone relation between evidence and the target. With help of these propositions we later derive how all explanations can be computed from a subset lattice.

We notice that some propositions for monotonicity in distribution are more restricting than similar propositions for monotonicity in mode. When constructing a method for computing all contrastive, counterfactual explanations, we want to impose the least amount of restrictions on the cases for which this method works. It follows that monotonicity in mode is more suitable for computing the explanations. As a result we proved some additional propositions for monotonicity in mode, for which no proofs were provided in case of monotonicity in distribution.

7.2.1 Monotonicity in mode and the sufficient set

Proposition 7.1. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t the highest ordered value for T and value e_i as the lowest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Proof. We prove this with contradiction.

Assume there is a sufficient set $\mathbf{S} \subseteq \mathbf{E}$ with observed value configuration $\mathbf{s}$ and $E_i \in \mathbf{S}$. Assume that E_i has observed value e_i which is the lowest ordered value for E_i. According to the definition $\mathbf{S}$ is a *minimal* set for which $\top(T|\mathbf{se}') = t$ holds for all value configurations for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. Let $\mathbf{S}'$ denote the set $\mathbf{S} \setminus \{E_i\}$. Because $\mathbf{S}$ is minimal, $\mathbf{S}'$ can not be a sufficient set as well. This means that there is a value e_i' for E_i where $\top(T|\mathbf{s}'\mathbf{e}'e_i') = t'$ with $\mathbf{s}'$ the observed value configuration for $\mathbf{S}'$ and $\mathbf{e}'$ an arbitrary value configuration for $\mathbf{E}'$. Let's check if this indeed holds. There are two cases; if E_i has an arbitrary value, it can take on value e_i or it can take on value e_i' with $e_i < e_i'$. In the first case, nothing is changed in comparison with the situation where E_i is included in sufficient set $\mathbf{S}$.

Now consider the second situation where $E_i = e'_i$. Because $e_i < e'_i$, we have $\mathbf{s'e'}e_i \preceq \mathbf{s'e'}e'_i$. Because E_i is isotone in mode, it follows that $\top(T|\mathbf{s'e'}e_i) \leq \top(T|\mathbf{s'e'}e'_i)$. t is the highest ordered value for T and $\top(T|\mathbf{s'e'}e_i) = t$, so the only possible value for $\top(T|\mathbf{s'e'}e'_i)$ is t as well.

We now have $\top(T|\mathbf{s'e''}) = t$ with $\mathbf{s'}$ the observed value configuration for $\mathbf{S}$ and for all value configurations $\mathbf{e''}$ for $\mathbf{E''} = \mathbf{E} \setminus \mathbf{S'}$. This leads to the contradiction that $\mathbf{S}$ is not minimal. So the observed value configuration for $\mathbf{S}$ is not a sufficient explanation. □

A similar proof can be given in case t is the lowest ordered value for T and e_i is the highest ordered value for E_i.

Proposition 7.2. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t the lowest ordered value for T and value e_i as the highest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Note that the previous propositions do not imply that an observation $E_i = e_i$ is necessarily part of a sufficient set for $T = t$ if e_i and t are both the highest or both the lowest ordered values for T and E_i and E_i is isotone in mode with T. Consider the following example. We have the observations $E_1 = e_1$ and $E_2 = e_2$ for variables $\{E_1, E_2\}$ that are both isotone in mode with target T. We also have $\top(T|e_1e_2) = t$. The possible values for E_1 are $e'_1 < e_1$ and and those for E_2 are $e'_2 < e_2$ with the following probability distribution;

$P(t|e_1e_2) = 0.9 \quad P(t|e_1e'_2) = 0.8$

$P(t|e'_1e_2) = 0.4 \quad P(t|e'_1e'_2) = 0.2$

The only sufficient set for $T = t$ is $\{E_1\}$. Both $E_2 = e_2$ and $T = t$ are the highest ordered values, however E_2 is not part of a sufficient set. On itself is the observation $E_2 = e_2$ not a sufficient set, because $\top(T|e'_1e_2) \neq t$. $\{E_1, E_2\}$ is however also not a sufficient set, because it is not minimal. So an observation $E_i = e_i$ is not necessarily part of a sufficient set for $T = t$ if e_i and t are both the highest or both the lowest ordered values for T and E_i and E_i is isotone in mode with T.

When evidence E_i is antitone in mode with T statements opposite to Proposition 7.1 and 7.2 are derived. The proofs for these propositions are analogous to the proof given for Proposition 7.1.

Proposition 7.3. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in mode with target T, $\top(T|\mathbf{e}) = t$ with t the highest ordered value for T and value e_i as the highest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Proposition 7.4. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in mode with target T, $\top(T|\mathbf{e}) = t$ with t the lowest ordered value for T and value e_i as the lowest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

In case the most probable value of the target is not its highest or lowest ordered value, we can still derive some information about the inclusion of a variable in a sufficient set.

Proposition 7.5. Given evidence $\mathbf{E}$, subset $\mathbf{S} \subset \mathbf{E}$ and $E_i \in \mathbf{E} \setminus \mathbf{S}$ isotone in mode with target T, $\top(T|\mathbf{e}) = t$, $\mathbf{S}$ and none of its subsets is sufficient and observed value e_i the lowest ordered value for E_i, $\mathbf{S} \cup \{E_i\}$ can only be a sufficient set if $\top(T|\mathbf{se'}) = t''$ with $t \leq t''$ for all value configurations $\mathbf{e'}$ of $\mathbf{E'} = \mathbf{E} \setminus \mathbf{S}$ not consistent with e_i.

Proof. We prove this with contradiction. Assume $\top(T|\mathbf{se'}) = t''$ with $t'' < t$. The lowest ordered value e_i of E_i is observed, so for all unobserved values e'_i for E_i we have $e_i < e'_i$. It follows that E_i has a value higher than observed in all value configurations $\mathbf{e'}$ not consistent with e_i. Let $\mathbf{e''}$ denote the value configuration for the set $\mathbf{E'} \setminus \{E_i\}$ that is consistent with $\mathbf{e'}$. We have $\mathbf{se''}e_i \preceq \mathbf{se'}$. Because E_i is isotone in mode with T it follows that $\top(T|\mathbf{se''}e_i) \leq \top(T|\mathbf{se'})$. Because $t'' < t$, $\mathbf{S} \cup \{E_i\}$ can not be a sufficient set. If we had instead assumed that $t \leq t''$, adding E_i to $\mathbf{S}$ could be a sufficient set. So we proved our statement. □

An analogous proof can be given in case the highest ordered value for E_i is observed. If E_i would be antitone in mode, the opposite propositions can be proven.

Proposition 7.6. Given evidence $\mathbf{E}$, subset $\mathbf{S} \subset \mathbf{E}$ and $E_i \in \mathbf{E} \setminus \mathbf{S}$ isotone in mode with target T, $\top(T|\mathbf{e}) = t$, $\mathbf{S}$ and none of its subsets is sufficient and observed value e_i the highest ordered value for E_i, $\mathbf{S} \cup \{E_i\}$ can only be a sufficient set if $\top(T|\mathbf{se'}) = t''$ with $t'' \leq t$ for all value configurations $\mathbf{e'}$ of $\mathbf{E'} = \mathbf{E} \setminus \mathbf{S}$ not consistent with e_i.

Proposition 7.7. Given evidence $\mathbf{E}$, subset $\mathbf{S} \subset \mathbf{E}$ and $E_i \in \mathbf{E} \setminus \mathbf{S}$ antitone in mode with target T, $\top(T|\mathbf{e}) = t$, $\mathbf{S}$ and none of its subsets is sufficient and observed value e_i as the lowest ordered value for E_i, $\mathbf{S} \cup \{E_i\}$ can only be a sufficient set if $\top(T|\mathbf{se'}) = t''$ with $t'' \leq t$ for all value configurations $\mathbf{e'}$ of $\mathbf{E'} = \mathbf{E} \setminus \mathbf{S}$ not consistent with e_i.

Proposition 7.8. Given evidence $\mathbf{E}$, subset $\mathbf{S} \subset \mathbf{E}$ and $E_i \in \mathbf{E} \setminus \mathbf{S}$ antitone in mode with target T, $\top(T|\mathbf{e}) = t$, $\mathbf{S}$ and none of its subsets is sufficient and observed value e_i as the highest ordered value for E_i, $\mathbf{S} \cup \{E_i\}$ can only be a sufficient set if $\top(T|\mathbf{se}') = t''$ with $t \leq t''$ for all value configurations $\mathbf{e}'$ of $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$ not consistent with e_i.

7.2.2 Monotonicity in distribution and the sufficient set

Similar propositions as the first four given in Section 7.2.1 can be given in case the evidence is isotone in *distribution* instead of isotone in mode with the target. However we have to assume that the target is binary valued. Similar propositions as 7.5 to 7.8 are not given when the evidence is monotone in distribution.

Proposition 7.9. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$ with t the highest ordered value for T and value e_i as the lowest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Proof. We prove this with contradiction. Assume there is a sufficient set $\mathbf{S} \subseteq \mathbf{E}$ with observed value configuration $\mathbf{s}$ and $E_i \in \mathbf{S}$.

According to the definition $\mathbf{S}$ is a *minimal* set with $\top(T|\mathbf{se}')$ for all value configurations for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. Let $\mathbf{S}'$ denote the set $\mathbf{S} \setminus \{E_i\}$. Because $\mathbf{S}$ is minimal, $\mathbf{S}'$ can not be a sufficient set as well. This means that there is a value configuration e_i' for E_i where $\top(T|\mathbf{s}'\mathbf{e}'e_i') = t'$ with $\mathbf{s}'$ the observed value configuration for $\mathbf{S}'$ and $\mathbf{e}'$ an arbitrary value configuration for $\mathbf{E}'$. Let's check if this indeed holds. There are two cases; if E_i has an arbitrary value, it can take on value e_i or it can take on value e_i' with $e_i < e_i'$. In the first case, nothing is changed in comparison with the situation where E_i is included in sufficient set $\mathbf{S}$.

Now consider the second situation, where $E_i = e_i'$. Because $e_i < e_i'$, we know that $\mathbf{s}'\mathbf{e}'e_i \preceq \mathbf{s}'\mathbf{e}'e_i'$. Because E_i is isotone in distribution, it follows that $P(T|\mathbf{s}'\mathbf{e}'e_i')$ is stochastically dominant over $P(T|\mathbf{s}'\mathbf{e}'e_i)$, which means that $P(T \leq t|\mathbf{s}'\mathbf{e}'e_i') \leq P(T \leq t|\mathbf{s}'\mathbf{e}'e_i)$ for all possible values of t of T. Because t' is the lowest value for T, we have that $P(t'|\mathbf{s}'\mathbf{e}'e_i') \leq P(t'|\mathbf{s}'\mathbf{e}'e_i)$. Since T is binary valued, it follows that $P(t|\mathbf{s}'\mathbf{e}'e_i') \geq P(t|\mathbf{s}'\mathbf{e}'e_i)$. We now have that $\top(T|\mathbf{s}'\mathbf{e}'e_i) = t$ and T is binary, so $P(t|\mathbf{s}'\mathbf{e}'e_i) > 0.5$. It now follows that $P(t|\mathbf{s}'\mathbf{e}'e_i') > 0.5$ as well.

We now have $\top(T|\mathbf{s}'\mathbf{e}'')$ given all value configurations $\mathbf{e}''$ for $\mathbf{E}'' = \mathbf{E} \setminus \mathbf{S}'$. This leads to the contradiction that $\mathbf{S}$ is not minimal and can thus not be a sufficient set. □

Again a similar proof can be given if t is the lowest ordered value and e_i the highest ordered value for T and E_i respectively. If we assume E_i is antitone in distribution with the target, the opposite can be deduced.

Proposition 7.10. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$ with t the lowest ordered value for T and value e_i as the highest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Proposition 7.11. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$ with t the highest ordered value for T and value e_i as the highest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

Proposition 7.12. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$ with t the lowest ordered value for T and value e_i as the lowest ordered value for E_i, observation $E_i = e_i$ can never be part of a sufficient explanation.

7.2.3 Monotonicity in mode and the counterfactual set

Proposition 7.13. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the lowest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i < e_i'$ can never be part of a counterfactual explanation.

Proof. We prove this with contradiction. Assume there is a counterfactual set $\mathbf{C} \subseteq \mathbf{E}$ and $E_i \in \mathbf{C}$ has observed value e_i. E_i takes on value e_i' with $e_i < e_i'$ in counterfactual explanation $\mathbf{c}$ corresponding to $\mathbf{C}$.

Because $\mathbf{c}$ is a counterfactual explanation, we have that $\top(T|\mathbf{ce}') = t'$ with $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$. It follows from the definition of the counterfactual set that there is no value configuration $\mathbf{c}'$ consistent with $\mathbf{c}$ for a subset $\mathbf{C}' \subset \mathbf{C}$ with $\top(T|\mathbf{c}'\mathbf{e}'') = t'$ with $\mathbf{e}''$ the observed value configuration for $\mathbf{E}'' = \mathbf{E} \setminus \mathbf{C}'$. Let $\mathbf{c}'$ denote the value configuration for $\mathbf{C}' = \mathbf{C} \setminus \{E_i\}$ consistent with $\mathbf{c}$ and let's check if this indeed holds.

It follows from $e_i < e_i'$, that $\mathbf{c}'\mathbf{e}'e_i \preceq \mathbf{ce}'$. Because E_i is isotone in mode with T, it follows that $\top(T|\mathbf{c}'\mathbf{e}'e_i) \leq \top(T|\mathbf{ce}')$. Because $\top(T|\mathbf{ce}') = t'$ and t' is the lowest ordered value for T, it follows that $\top(T|\mathbf{c}'\mathbf{e}'e_i) = t'$. This means there is value configuration $\mathbf{c}'$ consistent with $\mathbf{c}$ for a subset $\mathbf{C}' \subset \mathbf{C}$ with $\top(T|\mathbf{c}'\mathbf{e}'') = t'$ with $\mathbf{e}''$ the observed value configuration for $\mathbf{E}''$. This leads to the contradiction that $\mathbf{c}$ is not a counterfactual explanation. □

An analogous proof can be given if we had assumed that t was the highest ordered value for T and $e_i' < e_i$. If we had assumed that E_i is antitone in mode with the target a similar proof can be given for the opposite statement. This gives us the following propositions;

Proposition 7.14. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the highest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i' < e_i$ can never be part of a counterfactual explanation.

Proposition 7.15. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the lowest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i' < e_i$ can never be part of a counterfactual explanation.

Proposition 7.16. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the highest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i < e_i'$ can never be part of a counterfactual explanation.

In case the expected value does not have the highest or the lowest value for T we can still derive something about the inclusion of E_i in a sufficient set if the highest or lowest ordered value for E_i is observed.

Proposition 7.17. Given evidence $\mathbf{E}$ and target T with expected value t', $E_i \in \mathbf{E}$ for which lowest ordered value e_i is observed and that is isotone in mode with T, and set $\mathbf{C} \subseteq \mathbf{E}$ that is not a counterfactual set, a configuration for $\mathbf{C} \cup \{E_i\}$ can only be a counterfactual explanation if $\top(T|\mathbf{ce}') = t''$ and $t'' \leq t'$, given some unobserved value configuration $\mathbf{c}$ for $\mathbf{C}$ and $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

Proof. We prove this with contradiction. Assume $\top(T|\mathbf{ce}') = t''$ and $t' < t''$. The lowest ordered value e_i of E_i is observed, it follows that for all values e_i' for E_i that can be part of a counterfactual explanation we have $e_i < e_i'$. Let $\mathbf{e}''$ denote the observed value configuration for the set $\mathbf{E}' \setminus \{E_i\}$. E_i has the observed value in $\mathbf{E}'$, so $\mathbf{ce}' \preceq \mathbf{ce}''e_i'$. Because E_i is isotone in mode it follows that $\top(T|\mathbf{ce}') \leq \top(T|\mathbf{ce}''e_i')$. Because $t' < t''$, the addition of an unobserved value for E_i to value configuration $\mathbf{c}$ is not a counterfactual explanation.

If we had instead assumed that $t'' \leq t'$, adding an unobserved value for E_i to $\mathbf{c}$ could be a counterfactual explanation. So we proved our statement. □

Proposition 7.18. Given evidence $\mathbf{E}$ and target T with expected value t', $E_i \in \mathbf{E}$ for which highest ordered value e_i is observed and that is isotone in mode with T, and set $\mathbf{C} \subseteq \mathbf{E}$ that is not a counterfactual set, a configuration for $\mathbf{C} \cup \{E_i\}$ can only be a counterfactual explanation if $\top(T|\mathbf{ce}') = t''$ and $t' \leq t''$, given some unobserved value configuration $\mathbf{c}$ for $\mathbf{C}$ and $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

Proposition 7.19. Given evidence $\mathbf{E}$ and target T with expected value t', $E_i \in \mathbf{E}$ for which lowest ordered value e_i is observed and that is antitone in mode with T, and set $\mathbf{C} \subseteq \mathbf{E}$ that is not a counterfactual set, a configuration for $\mathbf{C} \cup \{E_i\}$ can only be a counterfactual explanation if $\top(T|\mathbf{ce}') = t''$ and $t' \leq t''$, given some unobserved value configuration $\mathbf{c}$ for $\mathbf{C}$ and $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

Proposition 7.20. Given evidence $\mathbf{E}$ and target T with expected value t', $E_i \in \mathbf{E}$ for which highest ordered value e_i is observed and that is antitone in mode with T, and set $\mathbf{C} \subseteq \mathbf{E}$ that is not a counterfactual set, a configuration for $\mathbf{C} \cup \{E_i\}$ can only be a counterfactual explanation if $\top(T|\mathbf{ce}') = t''$ and $t'' \leq t'$, given some unobserved value configuration $\mathbf{c}$ for $\mathbf{C}$ and $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

7.2.4 Monotonicity in distribution and the counterfactual set

Similar proposition as Propositions 7.13 to 7.16 can be given if the evidence in monotone in distribution with the target and the target is binary valued.

Proposition 7.21. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the lowest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i < e_i'$ can never be part of a counterfactual explanation.

Proof. We again prove this with contradiction. $E_i \in \mathbf{E}$ has observed value e_i and is isotone in mode with T. Assume there is a counterfactual set $\mathbf{C} \subseteq \mathbf{E}$ with configuration $\mathbf{c}$ containing $E_i = e_i'$ with $e_i < e_i'$. Because $\mathbf{c}$ is a counterfactual explanation and T is binary, we have $P(t'|\mathbf{ce}') > 0.5$ with $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$. It follows from the definition of a counterfactual set, that there is no value configuration $\mathbf{c}'$ consistent

with $\mathbf{c}$ for a subset $\mathbf{C}' \subset \mathbf{C}$ with $\top(T|\mathbf{c}'\mathbf{e}'') = t'$ with $\mathbf{e}''$ the observed value configuration for $\mathbf{E}'' = \mathbf{E} \setminus \mathbf{C}'$. Let $\mathbf{c}'$ denote the value configuration for $\mathbf{C}' = \mathbf{C} \setminus \{E_i\}$ consistent with $\mathbf{c}$ and let's check if this is indeed the case.

It follows from $e_i < e_i'$, that $\mathbf{c}'\mathbf{e}'e_i \preceq \mathbf{c}\mathbf{e}'$. Because E_i is isotone in distribution with T, we have that $P(T|\mathbf{c}\mathbf{e}')$ is stochastically dominant over $P(T|\mathbf{c}'\mathbf{e}'e_i)$, so $P(T \leq t|\mathbf{c}\mathbf{e}') \leq P(T \leq t|\mathbf{c}'\mathbf{e}'e_i)$ for all values t of T. Because t' is the lowest ordered value, we have that $P(t'|\mathbf{c}\mathbf{e}') \leq P(t'|\mathbf{c}'\mathbf{e}'e_i)$. Because $P(t'|\mathbf{c}\mathbf{e}') > 0.5$, it follows that $P(t'|\mathbf{c}'\mathbf{e}'e_i) > 0.5$ as well, so t' is the most probable value of T given $\mathbf{c}'\mathbf{e}'e_i$. This means there is a value configuration $\mathbf{c}'$ consistent with $\mathbf{c}$ for a subset $\mathbf{C}' \subset \mathbf{C}$ with $\top(T|\mathbf{c}'\mathbf{e}'') = t'$ with $\mathbf{e}''$ the observed value configuration for $\mathbf{E}''$. This leads to the contradiction that $\mathbf{c}$ is not a counterfactual explanation. □

A similar proof can be given for the case where t' is the highest ordered value for T and $e_i' < e_i$. Again the opposite is true for evidence variables that are antitone in distribution with the target. This leads to the following propositions.

Proposition 7.22. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is isotone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the highest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i' < e_i$ can never be part of a counterfactual explanation.

Proposition 7.23. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the lowest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i' < e_i$ can never be part of a counterfactual explanation.

Proposition 7.24. Given evidence $\mathbf{E}$, such that $E_i \in \mathbf{E}$ is antitone in distribution with binary valued target T, $\top(T|\mathbf{e}) = t$, expected value $t' \neq t$ for T with t' the highest ordered value for T and observed value e_i for E_i, a value assignment $E_i = e_i'$ with $e_i < e_i'$ can never be part of a counterfactual explanation.

8 Using Monotonicity to compute the explanation

If all evidence in a network has a monotone relation with the target, we can use this knowledge to more efficiently compute the explanation. In the previous section two different types of monotonicity were given; monotonicity in mode and in distribution. Some propositions were proven for both kinds of monotonicity. However, for some propositions the additional assumption of a binary valued target was needed to prove the same statement for evidence monotone in distribution as for evidence monotone in mode. Because we do not want to restrict our algorithm of computing all explanations to only binary valued targets, we only define how monotonicity in mode can be used to compute all explanations.

In Section 8.1 we derive how the propositions given in the previous section can be used to exclude certain evidence variables from the lattice. This gives a smaller lattice, so the explanations can be more efficiently computed. In Section 8.2 we give a definition of the monotonicity enhanced subset lattice. In Section 8.3 and 8.4 we discuss how the sufficient and counterfactual explanations can be derived from this lattice. We show how this works in practice in Section 8.5. In Section 8.6 we conclude with some observations about the given algorithms.

In the following sections we again assume that the BN is pruned before the explanations are computed.

8.1 Excluding evidence variables from the lattice

In Section 7, we presented multiple propositions about the relation between monotonicity and the inclusion of variables in the sufficient or counterfactual set. We proved that an evidence variable could not be part of the sufficient or counterfactual explanation based on the place in the ordering of the observed value and the ordering of the expected and most probable value of the target. If an evidence variable can not be part of the whole explanation, we do not have to take it into account during the computations. We decide to exclude this variable from the subset lattice. In this way no unnecessary probabilities are computed, which makes the process of finding all explanations more efficient.

Propositions 7.1, 7.2, 7.3 and 7.4 state when an evidence variable is not part of a sufficient explanation. Propositions 7.13, 7.14, 7.15 and 7.16 state when an evidence variable is not part of a counterfactual explanation. If it is derived from one of these propositions that an evidence variable is part of neither a sufficient nor a counterfactual set, we can exclude that variable from all subsets that are explored to find all explanations.

However, there are also situations where we know an evidence variable is not in any sufficient set, but we are not able to derive that it also is not part of any counterfactual sets or vice versa. In those situations we still exclude the evidence from the lattice. This gives the additional task of keeping track of all variables that are not in the lattice,

but could be part of the sufficient or counterfactual sets. We also need to compute some additional probabilities in the lattice to be sure all sufficient and counterfactual sets are still found. However, the benefit of having a lattice of reduced size outweighs these additional costs.

Given evidence E_i that is isotone in mode with the target, Table 2 gives an overview of when it can be derived that E_i is not part of a sufficient or a counterfactual set for different values of t, t' and the observed value e_i of E_i. In Table 3 the same is given for a piece of evidence that is antitone in mode with the target. Take the sixth row in Table 2 as an example. Given evidence $\mathbf{E}$, the target has most probable value t and expected value t' with t its highest ordered value and $t' < t$. We have $E_i \in \mathbf{E}$ that is isotone in mode with T and whose lowest ordered value is observed. We derive from Proposition 7.1 that E_i is not part of any sufficient set, because its lowest ordered value is observed. However, we do not have enough information to derive anything about its inclusion in a counterfactual set. The conclusion is that we exclude E_i while the lattice is built, but we should make some additional computations to find all counterfactual sets. We have that the lowest ordered value e_i for E_i is observed, so it follows from Proposition 7.17 that adding E_i to a set $\mathbf{C}$ can only give a counterfactual set if $\top(T|\mathbf{ce}') = t''$ with $t'' < t'$ for unobserved value configuration $\mathbf{c}$ and the observed values for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$. It follows that we only have to compute extra probabilities for sets in the lattice for which the previous statement holds when E_i is removed from the lattice. In all other situations where evidence is removed from the lattice but could still be included in a sufficient or counterfactual set, similar derivations could be made to decide when extra computations needs to be done.

In some cases we do exclude evidence even though we are not able to conclude anything about its inclusion in a sufficient or counterfactual set. Take for example the eighth row in Table 2. We have E_i that is isotone in mode with the target and E_i's lowest ordered value is observed. We also have an expected value that is lower ordered than the most probable value for T. Propositions 7.1 to 7.4 or Propositions 7.13 to 7.16 can not be used to derive that E_i is not part of any sufficient or counterfactual set. However, E_i is still excluded from the lattice. The resulting lattice as defined in the following section will have some useful properties if E_i is excluded from it. If E_i is included, the lattice does not have these properties which makes it harder to compute all explanations.

All evidence variables that are excluded from the lattice are placed in a set $\mathbf{X}$, $\mathbf{R}$ or $\mathbf{D}$ according to the reason why it was removed. $\mathbf{X}$ contains all evidence variables that can *not* be in any sufficient or counterfactual set. We use $\mathbf{R}$ to denote all evidence variables that are excluded from the lattice, but could still be in a sufficient set. $\mathbf{D}$ is used to denote all evidence variables that are excluded from the lattice, but could still be in a counterfactual set. These sets are formally defined as follows.

Definition 8.1. Given a set of evidence $\mathbf{E}$ that is monotone in mode with target T with most probable value t and expected value t' and $t \neq t'$, $\mathbf{X}$ includes all evidence variables $E_i \in \mathbf{E}$ where one of the following conditions hold:

- E_i is isotone in mode and t, t' and the observed value for E_i have an ordering as given by row 5 or 9 of Table 2.
- E_i is antitone in mode and t, t' and the observed value for E_i have an ordering as given by row 1 or 13 of Table 3.

Definition 8.2. Given a set of evidence $\mathbf{E}$ that is monotone in mode with target T with most probable value t and expected value t' and $t \neq t'$, $\mathbf{R}$ includes all evidence variables $E_i \in \mathbf{E}$ where one of the following conditions hold:

- E_i is isotone in mode and t, t' and the observed value for E_i have an ordering as given by row 7, 8, 11 or 12 of Table 2.
- E_i is antitone in mode and t, t' and the observed value for E_i have an ordering as given by row 3, 4, 15 or 16 of Table 3.

Definition 8.3. Given a set of evidence $\mathbf{E}$ that is monotone in mode with target T with most probable value t and expected value t' and $t \neq t'$, $\mathbf{D}$ includes all evidence variables $E_i \in \mathbf{E}$ where one of the following conditions hold:

- E_i is isotone in mode and t, t' and the observed value for E_i have an ordering as given by row 6, 8, 10 or 12 of Table 2.
- E_i is antitone in mode and t, t' and the observed value for E_i have an ordering as given by row 2, 4, 14 or 16 of Table 3.

From these definitions it follows that $\mathbf{X} \cap \mathbf{R} = \emptyset$, $\mathbf{X} \cap \mathbf{D} = \emptyset$. However we could have evidence that is both in $\mathbf{D}$ and in $\mathbf{R}$. We use $\mathbf{E^L}$ to denote the subset of $\mathbf{E}$ that is used to build the lattice.

Definition 8.4. Given a set of evidence $\mathbf{E}$ that is monotone in mode with target T with most probable value t and expected value t', $t \neq t'$ and set $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{D}$ according to Definition 8.1, 8.2 and 8.3, $\mathbf{E^L}$ denotes the set $\mathbf{E} \setminus (\mathbf{X} \cup \mathbf{R} \cup \mathbf{D})$.

Note that $\mathbf{E}$ still denotes the whole evidence set that was used to compute the most probable value for T in the BN, so $\mathbf{E} = \mathbf{E^L} \cup \mathbf{X} \cup \mathbf{R} \cup \mathbf{D}$. In the following sections we explain how the lattice is defined and how the explanations can be computed from it. It will be explained in those sections how the evidence in $\mathbf{X}, \mathbf{R}$ and $\mathbf{D}$ is treated to find all explanations.

8.2 Monotonicity enhanced subset lattice

Just as in the situation where all evidence was binary valued, we want to give each node in the lattice a label that can be used to determine if a subset is sufficient or counterfactual. In case of the binary valued evidence, a node $\mathbf{S}$ in the lattice is labeled with the most probable value of the target given the observed values for $\mathbf{S}$ and the unobserved values for $\mathbf{E} \setminus \mathbf{S}$. If the evidence is not binary valued, there are multiple unobserved value configurations for $\mathbf{E} \setminus \mathbf{S}$, which makes it less straightforward how the labels in the lattice need to be computed. We want to derive two things from a label; Firstly, we want to be able to derive if t is the most probable value given the observed values for $\mathbf{S}$ and all value configurations for $\mathbf{E} \setminus \mathbf{S}$. We can use this information to derive if $\mathbf{S}$ is a sufficient set. To accomplish this, we want to assign $\mathbf{E} \setminus \mathbf{S}$ a value configuration that makes value t the least likely. Because we have a monotonicity relation between the evidence and the target, we are able to derive which value configuration this is. If this monotonicity relation would not be present, we would not be able to derive this. Secondly, we want to derive from a label if the inverse of the set in the lattice is a counterfactual set. To do this, the evidence in the inverse of a node in the lattice should take on the value configuration that makes the *expected* value of the target the most likely. Again, the monotonicity relation between the evidence and the target can be used to derive what value configuration this is.

Even though the evidence variables in $\mathbf{X}, \mathbf{R}$ and $\mathbf{D}$ are excluded from the lattice, they still need to take on a value when the labels in the lattice are computed. We note that the observed values for those variables are exactly those values that make the most probable value of the target the least likely or the expected value the most likely.

This gives us the following definition of the monotonicity enhanced subset lattice.

Definition 8.5. (Monotonicity enhanced subset lattice) Given a target T and evidence $\mathbf{E}$ that is monotone in mode with T, most probable value t and expected value t' of T with $t \neq t'$ and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ according to Definitions 8.1, 8.2, 8.3 and 8.4, the monotonicity enhanced subset lattice is the subset lattice of $\mathbf{E^L}$ where each subset $\mathbf{S} \subseteq \mathbf{E^L}$ is labeled with the most probable value of T given the observed values for $\mathbf{S}$, $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{D}$ and the following values for all $E_i \in \mathbf{E^L} \setminus \mathbf{S}$:

- If $t' < t$ and E_i is isotone in mode with T, E_i takes on its lowest ordered value.
- If $t' < t$ and E_i is antitone in mode with T, E_i takes on its highest ordered value.
- If $t < t'$ and E_i is isotone in mode with T, E_i takes on its highest ordered value.
- If $t < t'$ and E_i is antitone in mode with T, E_i takes on its lowest ordered value.

Given a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, we call $\mathbf{C}$ with $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ the inverse of $\mathbf{S}$.

With this definition and the evidence variables that were excluded from the lattice it is enforced that all evidence variables in $\mathbf{E^L} \setminus \mathbf{S}$ always take on a value that is different than its observed value when the label for $\mathbf{S} \subseteq \mathbf{E^L}$ is computed. All evidence variables in $\mathbf{S} \cup \mathbf{X} \cup \mathbf{R} \cup \mathbf{D}$ take on their observed values when the label for $\mathbf{S}$ is computed. With these values we are able to derive if $\mathbf{S}$ is a (potential) sufficient set and if $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ is a (potential) counterfactual set. How this is done is explained in the following sections.

To be able to prove the correctness of our methods of deriving the sufficient and counterfactual sets from the lattice, we give the following propositions about the labels in the monotonicity enhanced subset lattice.

Proposition 8.1. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ and expected value t' with $t' < t$, and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, a subset in the lattice given by Definition 8.5 has a label that is equal to or lower ordered than the label of each of its parents.

$E_i \in \mathbf{E}$ is isotone in mode with T											
	t	t'	Order	e_i	$E_i \in \mathbf{S}$	Prop	$E_i \in \mathbf{C}$	Prop	Excluded	Set	Prop
1	H	L	$t' < t$	H	?		?		No		
2	H	–	$t' < t$	H	?		?		No		
3	–	L	$t' < t$	H	?		?		No		
4	–	–	$t' < t$	H	?		?		No		
5	H	L	$t' < t$	L	No	7.1	No	7.13	Yes	**X**	
6	H	–	$t' < t$	L	No	7.1	?		Yes, additional computations necessary to find all counterfactual sets	**D**	7.17
7	–	L	$t' < t$	L	?		No	7.13	Yes, additional computations necessary to find all sufficient sets	**R**	7.5
8	–	–	$t' < t$	L	?		?		Yes, additional computations necessary to find all sufficient and counterfactual sets	**D**, **R**	7.17, 7.5
9	L	H	$t < t'$	H	No	7.2	No	7.14	Yes	**X**	
10	L	–	$t < t'$	H	No	7.2	?		Yes, additional computations necessary to find all counterfactual sets	**D**	7.18
11	–	H	$t < t'$	H	?		No	7.14	Yes, additional computations necessary to find all sufficient sets	**R**	7.6
12	–	–	$t < t'$	H	?		?		Yes, additional computations necessary to find all sufficient and counterfactual sets	**D**, **R**	7.18, 7.6
13	L	H	$t < t'$	L	?		?		No		
14	L	–	$t < t'$	L	?		?		No		
15	–	H	$t < t'$	L	?		?		No		
16	–	–	$t < t'$	L	?		?		No		

Table 2: Given evidence E_i that is isotone in mode with the target T, this table states when it can be derived that E_i can be excluded from the lattice based on the placement in the ordering of the observed value of E_i, and the placement in the ordering of most probable value t and expected value t'. H indicates that a value is the highest ordered value, L indicates that a value is the lowest ordered value and $-$ indicates that a value is neither the highest or the lowest ordered value. The appropriate propositions are given that were used to derive the values in the table. The propositions in the last column can be used to decide when E_i can be part of the sufficient or counterfactual set.

$E_i \in \mathbf{E}$ is antitone in mode with T											
	t	t'	Order	e_i	$E_i \in \mathbf{S}$	Prop	$E_i \in \mathbf{C}$	Prop	Excluded	Set	Prop
1	H	L	$t' < t$	H	No	7.3	No	7.15	Yes	**X**	
2	H	–	$t' < t$	H	No	7.3	?		Yes, additional computations necessary to find all counterfactual sets	**D**	7.20
3	–	L	$t' < t$	H	?		No	7.15	Yes, additional computations necessary to find all sufficient sets	**R**	7.8
4	–	–	$t' < t$	H	?		?		Yes, additional computations necessary to find all sufficient and counterfactual sets	**D**, **R**	7.20, 7.8
5	H	L	$t' < t$	L	?		?		No		
6	H	–	$t' < t$	L	?		?		No		
7	–	L	$t' < t$	L	?		?		No		
8	–	–	$t' < t$	L	?		?		No		
9	L	H	$t < t'$	H	?		?		No		
10	L	–	$t < t'$	H	?		?		No		
11	–	H	$t < t'$	H	?		?		No		
12	–	–	$t < t'$	H	?		?		No		
13		H	$t < t'$	L	No	7.4	No	7.16	Yes	**X**	
14	L	–	$t < t'$	L	No	7.4	?		Yes, additional computations necessary to find all counterfactual sets	**D**	7.19
15	–	H	$t < t'$	L	?		No	7.16	Yes, additional computations necessary to find all sufficient sets	**R**	7.7
16	–	–	$t < t'$	L	?		?		Yes, additional computations necessary to find all sufficient and counterfactual sets	**D**, **R**	7.19 7.7

Table 3: Given evidence E_i that is antitone in mode with the target T, this table states when it can be derived that E_i can be excluded from the lattice based on the placement in the ordering of the observed value of E_i, and the placement in the ordering of most probable value t and expected value t'. H indicates that a value is the highest ordered value, L indicates that a value is the lowest ordered value and $-$ indicates that a value is neither the highest or the lowest ordered value. The appropriate propositions are given that were used to derive the values in the table. The propositions in the last column can be used to decide when E_i can be part of the sufficient or counterfactual set.

Proof. Assume we have a subset $\mathbf{S}$ with parent $\mathbf{P}$ that is labeled with t, where t is the lowest ordered label for all parents of $\mathbf{S}$. Let $E_i \in \mathbf{P}$ with $E_i \notin \mathbf{S}$ be isotone in mode with T and let e_i be the observed value for E_i. Let $\mathbf{e}'$ denote the value configuration for $\mathbf{E} \setminus \{E_i\}$ that was used to compute the label for $\mathbf{P}$. It follows from Definition 8.5 that $\mathbf{e}'$ was also used to compute the label for $\mathbf{S}$.

When the label for $\mathbf{P}$ is computed, E_i takes on its observed value e_i. Because E_i is isotone in mode with T and $t' < t$, it follows from Definition 8.5, that E_i takes on its lowest ordered value when the label for $\mathbf{S}$ is computed. Let e_i' denote this lowest ordered value. We have $e_i \neq e_i'$, because it follows from Table 2 that E_i would not be part of the lattice if its lowest ordered value was observed. We now have $\mathbf{e}'e_i' \preceq \mathbf{e}'e_i$. Because E_i is isotone in mode with T, $\top(T|\mathbf{e}'e_i') \leq \top(T|\mathbf{e}'e_i)$. It follows that the label of $\mathbf{S}$ is equal to or lower ordered than the label of $\mathbf{P}$.

If we have assumed E_i was antitone in mode with T a similar proof can be given. It is concluded that $\mathbf{S}$ has a label that is equal to or lower ordered than $\mathbf{P}$. Because it was assumed that $\mathbf{P}$ has the lowest label of all parents of $\mathbf{S}$, the label of $\mathbf{S}$ is equal to or lower than the labels of all parents of $\mathbf{S}$. □

When the expected value for T is higher ordered than the most probable value, a similar proof can be given to show that the label of a node is always equal to or higher ordered than the labels of its parents. This leads to the following proposition.

Proposition 8.2. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ and expected value t' with $t < t'$, and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, a node in the lattice given by Definition 8.5 has a label that is equal to or higher ordered than the label of each of its parents.

8.3 Deriving sufficient explanations from the lattice

In this section we give an algorithm for finding all sufficient explanations based on different propositions. The resulting algorithm consists of two parts. The first part performs a breadth-first search through the monotonicity enhanced subset lattice to find all sufficient sets containing only evidence variables represented in the lattice. In case the set $\mathbf{R}$ with evidence variables excluded from the lattice is not empty, not all sufficient sets will be found by this first search. The second part of the algorithm revisits some specific nodes in the lattice that can be combined with evidence variables in $\mathbf{R}$ to give additional sufficient sets. The two parts combined give all possible sufficient sets. We first explain in Section 8.3.1 how the breadth-first search can be done. Afterward we show in Section 8.3.2 how we can find all remaining sufficient sets.

8.3.1 Finding sufficient explanations without the excluded evidence

We make a distinction between two different cases in finding all sufficient sets containing only evidence variables represented in the lattice. The first is the case where most probable value t given all evidence $\mathbf{E}$ is the highest or the lowest ordered value for target T. In this situation, only the labels in the lattice are needed to derive if a set is sufficient. The second is the case where t is not the highest or lowest ordered value for T. Now the labels in the lattice do not provide enough information to derive if a set is sufficient. An additional probability needs to be computed to decide this.

We first prove how the sufficient sets can be found in the lattice if t is the highest or lowest ordered value for T.

Proposition 8.3. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t either the highest or the lowest ordered value for T and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, if a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5 is labeled with t and none of its children is labeled with t, then $\mathbf{S}$ is a sufficient set.

Proof. It follows from Definition 8.2 that $\mathbf{R}$ is always empty if t is the highest or lowest ordered value for T. All sufficient sets can thus be found in the lattice.

Assume t is the highest ordered value for T and consider a subset $\mathbf{S} \subseteq \mathbf{E^L}$ that is labeled with t and none of its children is labeled with t.

According to Definition 8.5, the label t of subset $\mathbf{S}$ represents the configuration where the evidence variables in $\mathbf{S} \cup \mathbf{X} \cup \mathbf{D}$ take on their observed values and all evidence variables E_i in $\mathbf{E}' = \mathbf{E^L} \setminus \mathbf{S}$ takes on their lowest or highest ordered value depending on whether E_i is isotone or antitone in mode. Let $\mathbf{s}$, $\mathbf{x}$ and $\mathbf{d}$ denote the value configurations with the observed values for $\mathbf{S}$, $\mathbf{X}$ and $\mathbf{D}$ respectively. If t is the most probable value for T given the observed valued for $\mathbf{S}$ and all value configurations for $\mathbf{E} \setminus \mathbf{S}$ and no subsets of $\mathbf{S}$ are sufficient sets, then $\mathbf{S}$ is a sufficient set.

Let $E_1 \in \mathbf{E}'$ be isotone in mode with T. Because t is the highest ordered value for T and E_1 is isotone in mode, it follows from Definition 8.5 that the label of $\mathbf{S}$ represents a configuration where E_1 takes on its lowest ordered

value. Because $\mathbf{S}$ is labeled with t, t is the most probable value for T given $\mathbf{sxd}$, the lowest ordered value for E_1 and the values for evidence variables in $\mathbf{E}' \setminus \{E_1\}$ as given by Definition 8.5. Because E_1 is isotone in mode with T, only values t' with $t \leq t'$ can be the most probable value for T given $\mathbf{sxd}$, the values for evidence variables in $\mathbf{E}' \setminus \{E_1\}$ as given by Definition 8.5 and *all* unobserved values for E_1. Because t is the highest ordered value for T, t is the most probable value given $\mathbf{sxd}$, the values for $\mathbf{E}' \setminus \{E_1\}$ as defined by Definition 8.5 and *all* values for E_1.

Let $E_2 \in \mathbf{E}'$ be antitone in mode with T. It follows from Definition 8.5 that the label for $\mathbf{S}$ is computed for a configuration where E_2 has its highest ordered value. It follows from the label of $\mathbf{S}$, that t is the most probable value for T given $\mathbf{sxd}$, the values of evidence variables in $\mathbf{E}' \setminus \{E_2\}$ as given by Definition 8.5 and the highest ordered value for E_2. Because E_2 is antitone in mode with T, only values t' with $t \leq t'$ can be the most probable value for T given $\mathbf{sxd}$ and the values for evidence variables in $\mathbf{E}' \setminus E_2$ as given by Definition 8.5 when E_2 is assigned a lower ordered value. Because t is the highest ordered value for T, t is the most probable value given $\mathbf{sxd}$ and the values for evidence variables in $\mathbf{E}' \setminus \{E_2\}$ as given by Definition 8.5 and *all* values for E_2.

Now let $D_1 \in \mathbf{X} \cup \mathbf{D}$ be isotone in mode with T. It follows from Definition 8.3 and 8.1 that the lowest ordered value for D_1 is observed. It follows from Definition 8.5 that this value was used to compute the label for $\mathbf{S}$. With the same logic as given for evidence variable E_1 it can be concluded that t is the most probable value given $\mathbf{s}$, the observed values for $\mathbf{X} \cup \mathbf{D} \setminus \{D_1\}$, the values for evidence variables in $\mathbf{E}'$ as given by Definition 8.5 and *all* values for D_1.

Let $D_2 \in \mathbf{X} \cup \mathbf{D}$ be antitone in mode with T. We can derive from Definition 8.3 and 8.1 that the highest ordered value for E_i is observed. It is given by Definition 8.5 that this value was used to compute the label for $\mathbf{S}$. The same derivation as for evidence E_2 can be made to conclude that t is the most probable value given $\mathbf{s}$, the observed values for $\mathbf{X} \cup \mathbf{D} \setminus \{D_2\}$, the values for evidence variables in $\mathbf{E}'$ as given by Definition 8.5 and *all* values for D_2.

We proved for all isotone and antitone evidence variables $E_i \in \mathbf{E}'$ that the most probable value t for T does not change if the value for E_i is changed when $\mathbf{s}$ is given and the values for all evidence variables in $\mathbf{E} \setminus (\mathbf{S} \cup \{\mathbf{E_i}\})$ are kept constant. The same is proven for all evidence variables in $\mathbf{X} \cup \mathbf{D}$. Because $\mathbf{E} \setminus \mathbf{S} = \mathbf{E}' \cup \mathbf{X} \cup \mathbf{D}$, it follows that t is the most probable value given the observed values for $\mathbf{S}$ and for all possible value configurations for $\mathbf{E} \setminus \mathbf{S}$. If $\mathbf{S}$ is a minimal set for which this is the case, $\mathbf{S}$ is a sufficient set.

We made the assumption that all children of $\mathbf{S}$ have a label different than t. It follows from Proposition 8.1 that these labels are lower ordered than t. As a consequence, all other descendants of $\mathbf{S}$ also have a label lower ordered than t. It follows that none of the descendants of $\mathbf{S}$ have label t, which means that they can not provide a sufficient set. It follows that the observed value configuration for $\mathbf{S}$ is a sufficient explanation.

If we assume t is the lowest ordered value for T, a similar proof can be given. It is concluded that the observed value configuration for a set $\mathbf{S}$ is a sufficient set if $\mathbf{S}$ is labeled with most probable value t in the lattice and all children of $\mathbf{S}$ have a different label. □

From the previous proposition we conclude that the sufficient sets can be found by only looking at the labels in the lattice, when the most probable value for the target is the lowest or the highest ordered value. In case most probable value t is not the highest or the lowest ordered value for T, we need some additional information to be able to conclude that a set is sufficient.

Consider the following situation. We have evidence $\mathbf{E}$ and target T. T has possible values $t' < t < t''$, with $\top(T|\mathbf{e}) = t$ and t' the expected value. Assume we have a set $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice that is labeled with t and no children of $\mathbf{S}$ in the lattice are labeled with t. Let $\mathbf{E}' = \mathbf{E^L} \setminus \mathbf{S}$ and say $E_i \in \mathbf{E}'$ is isotone in mode with T and has possible values $e_i' < e_i < e_i''$ with e_i observed. Because E_i is isotone in mode with T and $t' < t$, it follows from Definition 8.5 that the label t of $\mathbf{S}$ represents a configuration where E_i takes on value e_i'. For this value configuration and the observed values for $\mathbf{S}$, t is the most probable value. Because E_i is isotone in mode with T, the values e_i and e_i'' can have t or a higher ordered value for T as most probable value given observed values for $\mathbf{S}$ and values for $\mathbf{E} \setminus (\mathbf{S} \cup \{E_i\})$ as given by Definition 8.5. Because t is not the highest ordered value for T, we are not able to derive if $\mathbf{S}$ is sufficient from its label alone. We conclude that an additional probability needs to be computed in case a set is labeled with t in the lattice and t is not the highest or the lowest ordered value for T.

We prove how we can find all sufficient sets in the lattice in case t is not the highest or the lowest ordered value for T.

Proposition 8.4. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T\mathbf{e}) = t$, with t not the highest or the lowest ordered value for T, expected value t' for T with $t \neq t'$ and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5 is sufficient if all of the following statements hold.

- $\mathbf{S}$ is labeled with t.
- $\top(T|\mathbf{se}') = t$, with $\mathbf{s}$ the observed values for $\mathbf{S}$ and $\mathbf{e}'$ is the value configuration where all $E_i \in \mathbf{E} \setminus \mathbf{S}$ take on
  - the highest ordered value if E_i had the lowest ordered value when the label for $\mathbf{S}$ was computed.
  - the lowest ordered value if E_i had the highest ordered value when the label for $\mathbf{S}$ was computed.
- None of the children of $\mathbf{S}$ are labeled with t or the children labeled with t are not sufficient.

Proof. Because t is not the highest or lowest ordered value for T, it follows from Definition 8.1 that $\mathbf{X}$ is empty. Given is a set $\mathbf{S}$ for which all three statements hold and let $\mathbf{E}' = \mathbf{E^L} \setminus \mathbf{S}$. Consider the situation where expected value t' of T is lower ordered than t.

Let $E_i \in \mathbf{E} \setminus \mathbf{S}$ be isotone in mode with T. We have $\mathbf{E} \setminus \mathbf{S} = \mathbf{E}' \cup \mathbf{R} \cup \mathbf{D}$, so it follows that $E_i \in \mathbf{E}'$ or $E_i \in \mathbf{R} \cup \mathbf{D}$. If $E_i \in \mathbf{E}'$, it follows from Definition 8.5 that E_i has the lowest ordered value when the label t for $\mathbf{S}$ is computed in the lattice, because $t' < t$ and E_i is isotone in mode. If $E_i \in \mathbf{R} \cup \mathbf{D}$ it follows from Definition 8.2 and 8.3 that E_i was excluded from the lattice because its lowest ordered value was observed. This values is used when t for $\mathbf{S}$ was computed in the lattice. So it follows that isotone $E_i \in \mathbf{E} \setminus \mathbf{S}$ always takes on its lowest ordered value when the label for $\mathbf{S}$ is computed. Because E_i is isotone in mode with T, it follows that higher ordered values for E_i can only have a most probable value t'' for T with $t \leq t''$, when the observed values for $\mathbf{S}$ and the values for $\mathbf{E} \setminus (\mathbf{S} \cup \{E_i\})$ as defined by Definition 8.5 are given.

In value configuration $\mathbf{e}'$ in the second statement of this proposition, E_i has its highest ordered value. Given the observed values for $\mathbf{S}$ and $\mathbf{e}'$, the most probable value for T is t and E_i is isotone in mode, so it follows that a lower ordered value for E_i can only have a most probable value t'' for T with $t'' \leq t$, when the observed values for $\mathbf{S}$ are given and the evidence variables in $\mathbf{E} \setminus (\mathbf{S} \cup \{E_i\})$ take on the value configuration as defined by the second statement. It follows that changing the value for a piece of evidence that is isotone in mode does not affect the most probable value for T. An analogous proof can be given to derive that changing the value for an antitone piece of evidence does not affect the most probable value for T. It follows that t is the most probable value for T given the observations for $\mathbf{S}$ and all value configurations of $\mathbf{E} \setminus \mathbf{S}$.

All children of $\mathbf{S}$ have a label different than t or are not a sufficient set, which means that $\mathbf{S}$ is a minimal set for which this is the case. We conclude that $\mathbf{S}$ is a sufficient set.

If we assume that expected value t' was higher ordered than t, a similar proof can be given. □

With the use of Proposition 8.3 and 8.4 a breadth-first search can be constructed to find sufficient sets in the lattice. All sufficient sets that are found by this algorithm are valid, however not all sufficient sets will be found. The sufficient sets that are not found by the algorithm are those sets that contain evidence in $\mathbf{R}$. A description of the breadth-first search algorithm that is started at the top of the lattice is given by Algorithm 5. While the breadth-first search is performed the lattice is dynamically constructed. In case t is the lowest or highest ordered value for T it follows from Proposition 8.3 that it can be concluded if a set is sufficient after its own label and the labels for all its children are computed. If t is not the lowest or highest ordered value for T, the algorithm computes the additional mode as indicated by Proposition 8.4. We notice that the search could also be started at the bottom of the lattice, because it is not dependent on the labels of the parents if a set is sufficient. We chose to only give the search starting at the top, because we will see in the following sections that this be could more easily combined with a search for all counterfactual sets.

8.3.2 Computing the remaining sufficient explanations

After the breadth-first search is performed, not all sufficient sets are necessarily found, because in some cases it is possible to add evidence in $\mathbf{R}$ to a subset in the lattice to give a sufficient set. If the most probable value t is the highest or lowest value for T, we derive from Definition 8.2 that $\mathbf{R}$ is empty, so there is no evidence excluded from the lattice that still could be part of a sufficient set. In case $\mathbf{R}$ is empty, it follows that all sufficient sets are found by Algorithm 5. In that situation it is not necessary to perform the additional checks described in this section.

We first give an example of a situation where a subset $\mathbf{R}' \subseteq \mathbf{R}$ can be added to a subset $\mathbf{S}$ in the lattice that is not sufficient to create a sufficient set. Afterward we prove that this example is the only situation in which evidence in $\mathbf{R}$ can be added to $\mathbf{S}$ to create a sufficient set. Consider for example the following situation. We have

Algorithm 5: Computes all sufficient sets containing only evidence in $\mathbf{E^L}$ with a breadth-first search through the subset lattice

Input : Target T with most probable value t and expected value t',
evidence set $\mathbf{E}$ monotone in mode with T used to build the lattice
Output: All sufficient sets

```
SuffcientSets = ∅
Q = Queue containing E
while Q not empty do
    S = get first item in Q
    e' = value configuration for E \ S as given by Definition 8.5
    t'' = ⊤(T|se')
    if t'' == t and t is highest or lowest ordered value then
        Put all children of S in Q
    else if t'' == t and t is not highest or lowest ordered value then
        e'' = value configuration for E \ S as given by Proposition 8.4
        if ⊤(T|se'') = t then
            Put all children of S in Q
        end
    end
end
return All sets labeled with t for which all children labeled with t are no sufficient set
```

a target T with expected value t' its lowest ordered value and most probable value t with $t' < t$. Let $E_i \in \mathbf{R}$ be isotone in mode with T. We derive from Definition 8.2 that E_i was excluded from the lattice because its lowest ordered value was observed. Let $\mathbf{S}$ be a set that is labeled with t in the lattice, but that is not sufficient because the configuration in the second condition from Proposition 8.4 gave value t'' with $t < t''$ as most probable value for T. If no subsets of $\mathbf{S}$ are sufficient sets, it now follows from Proposition 7.5 that $\mathbf{S} \cup \{E_i\}$ could be a sufficient set. So in this situation we have to try for each subset $\mathbf{R'} \in \mathbf{R}$ if $\mathbf{S} \cup \mathbf{R'}$ is a sufficient set.

In the following two propositions we first prove in which situations adding a subset $\mathbf{R'} \in \mathbf{R}$ to a set $\mathbf{S}$ in the lattice potentially gives a sufficient set. Afterward we prove what mode needs to be computed to verify whether $\mathbf{S} \cup \mathbf{R'}$ indeed is a sufficient set.

Proposition 8.5. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t not the highest or the lowest ordered value for T, expected value t' for T with $t' < t$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, and a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, $\mathbf{S} \cup \mathbf{R'}$ with $\mathbf{R'} \subseteq \mathbf{R}$ can only be a sufficient set if the following statements hold.

- $\mathbf{S}$ is labeled with t in the lattice.
- The computation of the mode in the second statement of Proposition 8.4 gives a value t'' with $t < t''$.
- No subsets of $\mathbf{S} \cup \mathbf{R'}$ are sufficient.

Proof. Because t is not the highest or lowest ordered value for T, it follows from Definition 8.1 that $\mathbf{X}$ is empty. Given is expected value t' for T with $t' < t$ and a subset $\mathbf{S}$ for which all statements hold.

We conclude that $\mathbf{S}$ is not a sufficient set, because the computation in the second statement of Proposition 8.4 gives a value that is not equal to t. It follows from the third statement of the current proposition that no subsets of $\mathbf{S}$ are sufficient. We conclude that adding evidence to $\mathbf{S}$ potentially gives a sufficient set.

Let $R_i \in \mathbf{R}$ be isotone in mode with the target. When the label t for $\mathbf{S}$ is computed, it follows from Definition 8.5 that R_i is assigned its observed value. It follows from Definition 8.2 that this observed value is its lowest ordered value. We conclude that t is the most probable value for T given the observed values for $\mathbf{S} \cup \{R_i\}$ and the values for $\mathbf{E} \setminus (\mathbf{S} \cup \{R_i\})$ as given by Definition 8.5. For all evidence $E_i \in \mathbf{E} \setminus (\mathbf{S} \cup \{R_i\})$ that is isotone in mode with T we can either derive from Definition 8.2, 8.3 or 8.5 that its lowest ordered value was used when the label for $\mathbf{S}$ is computed. For all evidence $E_i \in \mathbf{E} \setminus (\mathbf{S} \cup \{R_i\})$ that is antitone in mode with T we can derive from the same definitions that its highest ordered value was used when the label for $\mathbf{S}$ is computed. It now follows from the monotonicity relation that a value t'' with $t \leq t''$ is the most probable value for T given all possible value configurations for $\mathbf{E} \setminus (\mathbf{S} \cup \{R_i\})$ and the observed values for $\mathbf{S} \cup \{R_i\}$. Because R_i is isotone in mode with T, it follows that a value t'' with $t \leq t''$

is the most probable value for T given all value configurations for $\mathbf{E} \setminus (\mathbf{S} \cup \{R_i\})$, the observed values for $\mathbf{S}$ and *all* values for R_i.

We conclude that a value t'' with $t \leq t''$ is the most probable value given $\mathbf{S}$, an unobserved value for R_i and all value configurations for $\mathbf{E} \setminus \mathbf{S}$. It now follows from Proposition 7.5 that $\mathbf{S} \cup \{R_i\}$ could give a sufficient set.

If we had assumed R_i was antitone in mode with the target a similar derivation could be made. In that case Proposition 7.8 is used to derive that $\mathbf{S} \cup \{R_i\}$ could give a sufficient set. It follows that for all subsets $\mathbf{S}$ in the lattice for which the three given statements hold, $\mathbf{S} \cup \mathbf{R}$ could be a sufficient set for all subsets $\mathbf{R} \subseteq \mathbf{R}$. □

It follows from the previous proposition for what subsets in the lattice adding $\mathbf{R}' \subseteq \mathbf{R}$ could provide a sufficient set in case the expected value is lower ordered than the most probable value for the target. In case the expected value is higher ordered than the most probable value, an analogous proof can be given for the following statement.

Proposition 8.6. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t not the highest or the lowest ordered value for T, expected value t' for T with $t < t'$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, and a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, $\mathbf{S} \cup \mathbf{R}'$ with $\mathbf{R}' \subseteq \mathbf{R}$ can only be a sufficient set if the following statements hold.

- $\mathbf{S}$ is labeled with t in the lattice.
- The computation of the mode in the second statement of Proposition 8.4 gives a value t'' with $t'' < t$.
- No subsets of $\mathbf{S} \cup \mathbf{R}'$ are sufficient.

If we have determined that $\mathbf{S} \cup \mathbf{R}'$ is potentially a sufficient set, one additional mode needs to be computed to verify this. We prove this with the following propositions. This proposition holds for all value orderings for t'.

Proposition 8.7. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ with t not the highest or the lowest ordered value for T, expected value t' for T with $t' \neq t$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, and a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, $\mathbf{S} \cup \mathbf{R}'$ with $\mathbf{R}' \subseteq \mathbf{R}$ is a sufficient set if the following statements hold.

- It follows from Proposition 8.5 or 8.6 that $\mathbf{S} \cup \mathbf{R}'$ could be a sufficient set.
- $\top(T|\mathbf{sr}'\mathbf{e}') = t$, with $\mathbf{sr}'$ the observed values for $\mathbf{S} \cup \mathbf{R}'$ and $\mathbf{e}'$ is the value configuration where all $E_i \in \mathbf{E} \setminus (\mathbf{S} \cup \mathbf{R}')$ take on
  - the highest ordered value if E_i had the lowest ordered value when the label for $\mathbf{S}$ was computed.
  - the lowest ordered value if E_i had the highest ordered value when the label for $\mathbf{S}$ was computed.

Proof. Because t is not the highest or lowest ordered value for T, it follows from Definition 8.1 that $\mathbf{X}$ is empty. Given is a set $\mathbf{S}$ in the lattice and a subset $\mathbf{R}' \subseteq \mathbf{R}$ for which all statements hold. Let $\mathbf{E}'$ denote the set $\mathbf{E} \setminus (\mathbf{S} \cup \mathbf{R}')$. Assume $t' < t$.

It follows from the proof given for Proposition 8.5 that a value t'' with $t \leq t''$ is the most probable value given $\mathbf{sr}'$ and all value configurations for $\mathbf{E}'$.

Let $\mathbf{e}'$ denote the value configuration for $\mathbf{E}'$ used to compute the mode in the second statement of this proposition. Let $E_i \in \mathbf{E}'$ be isotone in mode with T. We derive that E_i takes on its highest ordered value in $\mathbf{e}'$. Because E_i is isotone in mode with the target, it follows that a lower ordered value for E_i has a most probable value that is equal to or lower ordered than t given the values for $\mathbf{E}' \setminus \{E_i\}$ as given in the second statement of this proposition and the observed values for $\mathbf{S} \cup \mathbf{R}'$. We stated before that given $\mathbf{sr}'$, all value configurations for $\mathbf{E}'$ give a value t'' with $t \leq t''$ as most probable value for T. It follows that a piece of evidence in $\mathbf{E}'$ that is isotone in mode with T can have a arbitrary value and t would be the most probable value given $\mathbf{S} \cup \mathbf{R}'$.

The same derivation can be made to state that evidence in $\mathbf{E}'$ that is antitone in mode can be changed arbitrarily without affecting the most probable value t given $\mathbf{S} \cup \mathbf{R}'$. It follows that t is the most probable value for T given the observations for $\mathbf{S} \cup \mathbf{R}'$ and all possible value configurations $\mathbf{E} \setminus (\mathbf{S} \cup \mathbf{R}')$. Because no subsets of $\mathbf{S} \cup \mathbf{R}'$ are sufficient, it follows that $\mathbf{S} \cup \mathbf{R}'$ is sufficient.

In case we had assumed $t < t'$, a similar proof can be given. □

It follows from the previous propositions how the sufficient sets that are not yet found by the breadth-first from Algorithm 5 search can be found. Propositions 8.5 and 8.6 are used to select the subsets in the lattice to which $\mathbf{R}' \subseteq \mathbf{R}$ can be added to give a sufficient set based on the ordering of the expected and most probable value of

the target. If this selection is made Proposition 8.7 is used to determine if the constructed set is indeed sufficient. During the breadth-first search, we should keep track of all subsets $\mathbf{S}$ that could be combined with evidence in $\mathbf{R}$ to give a sufficient set according to Proposition 8.5 and 8.6. For all those subsets we loop over the subsets $\mathbf{R}' \subseteq \mathbf{R}$ to compute the mode given in Proposition 8.7 to decide if $\mathbf{S} \cup \mathbf{R}'$ is sufficient. During this loop we should use Propositions 7.5 to 7.8 and the minimality requirement of the sufficient set to terminate the search as early as possible. Algorithm 6 gives how this can be done in case the expected value is lower ordered than the most probable value for T. The same algorithm where $<$ is flipped can be given if the expected value is higher ordered than the most probable value for T.

Algorithm 6: Computes all sufficient sets with a breadth-first search through the subset lattice

Input : Target T with most probable value t and expected value t' the lowest ordered value for T,
set $\mathbf{R}$ monotone with T with evidence removed from the lattice,
$\mathbf{S}'$ with all sets in the lattice that could give a sufficient set according to Proposition 8.5
Output: All sufficient sets

```
SuffcientSets = ∅
foreach S ∈ S' do
    foreach R' ⊆ R do
        e' = value configuration for E \ (S ∪ R') as given by Definition 8.5
        t'' = ⊤(T|sr'e')
        if t'' == t then
            Add S ∪ R' to SuffcientSets
            break
        else if t'' < t then
            break
        end
    end
end
return SuffcientSets
```

8.4 Deriving counterfactual explanations from the lattice

Computing the counterfactual sets is less straightforward for evidence variables with multiple values than in case of binary valued evidence, because an evidence variable has multiple unobserved values. With binary valued evidence, if we found a node whose inverse is a counterfactual set, none of its descendants could give a counterfactual set, because they would not be minimal. This is however not the case with evidence that is not binary valued. Consider for example a situation where $E_1 = e_1'$ and $E_1 = e_1, E_2 = e_2$ are both counterfactual explanations for a target T. Even though $\{E_1\}$ is a subset of the variables in the second explanation, both are minimal counterfactual explanations because E_1 is assigned a different value in both explanations. Because $\{E_1\}$ is a subset of the other set, its inverse is an ancestor of the inverse of $\{E_1, E_2\}$ in the lattice. It follows that we will not find all counterfactual sets, when the children of subsets in the lattice labeled with expected target value t' are not explored.

In contrast to the algorithm for finding all sufficient sets, finding all counterfactual sets in the lattice can be done with one breadth-first search. We first describe in Section 8.4.1 a breadth-first search that finds all counterfactual sets containing only evidence variables represented in the lattice. During the search it is computed for the appropriate unobserved value configurations if they are counterfactual explanations. This breath-first search does not find counterfactual sets that also contain evidence variables not represented in the lattice. We explain in Section 8.4.2 how the algorithm can be extended to also find all counterfactual explanations with evidence variables excluded from the lattice.

8.4.1 Finding counterfactual explanations without the excluded evidence variables

To prove where we can find all counterfactual sets that only contain variables from the inverses of sets in the lattice, we make a distinction between a situation where expected value t' is the lowest or the highest ordered value and where t' is not the highest or lowest ordered value.

First we prove how the counterfactual sets can be derived from the lattice when t' is the highest or lowest ordered value for T. We prove that only subsets labeled with t' can give a counterfactual set in this situation.

Proposition 8.8. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value t', that is either the highest or the lowest ordered value for T and $t \neq t'$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, if a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5 is not labeled with t', then $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ is not a counterfactual set.

Proof. Because t' is the highest or lowest ordered value for T, it follows from Definition 8.3 that $\mathbf{D}$ is empty. Assume that t' is the lowest ordered value for T and assume we have a subset $\mathbf{S} \subseteq \mathbf{E^L}$ with its inverse $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$.

Let $E_i \in \mathbf{C}$ be isotone in mode with T. It follows from Definition 8.5 that E_i takes on its lowest ordered value when computing the label for $\mathbf{S}$. All evidence $\mathbf{E} \setminus \mathbf{C}$ takes on its observed value when the label for $\mathbf{S}$ is computed. Denote this label with $t'' \neq t'$. Because t' is the lowest ordered value, $t' < t''$. Because E_i is isotone in mode with T, a higher ordered value for E_i can only give a value equal to or higher ordered than t'' as most probable value for T with the observed values for $\mathbf{E} \setminus \mathbf{C}$ and the values for $\mathbf{C} \setminus \{E_i\}$ as given by Definition 8.5. Because t' is the lowest ordered value, it follows that no value configuration for an unobserved value for E_i and values for $\mathbf{C} \setminus \{E_i\}$ as given by Definition 8.5 is a counterfactual explanation. If we had assumed that E_i is antitone in mode, we can prove the same statement. If follows $\mathbf{C}$ is not a counterfactual set.

If we had assumed that t' is the highest ordered value for T, a similar proof can be given. □

It follows from the previous proposition that we only have to consider nodes labeled with t' in the lattice to find all counterfactual sets when t' is the lowest or highest ordered value for T. In case the expected value is not the highest or the lowest value, the inverse of a subset that is not labeled with t' in the lattice can provide a counterfactual set. We first prove where all potential counterfactual sets can be found if the expected value is lower than the most probable value of the target. Afterwards we prove how we can check if the corresponding unobserved value configurations are counterfactual explanations.

Proposition 8.9. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T\mathbf{e}) = t$ and expected value t' with $t' < t$ and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, if a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5 is labeled with t'' and $t'' < t'$, $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ is potentially a counterfactual set.

Proof. Assume we have a subset $\mathbf{S} \subseteq \mathbf{E^L}$ that is labeled with $t'' < t'$ and has inverse $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$. It follows from Definition 8.5 that all evidence variables in $\mathbf{E} \setminus \mathbf{C}$ take on their observed values when the label for $\mathbf{S}$ is computed. Let $E_i \in \mathbf{C}$ be isotone in mode. It follows from Definition 8.5 that E_i takes on its lowest ordered value when computing this label. Because E_i is isotone in mode with T, a higher ordered value for E_i can give a value equal to or higher ordered than t'' as most probable value for T given the observed values for $\mathbf{E} \setminus \mathbf{C}$ and the values for $\mathbf{C} \setminus \{E_i\}$ as given by Definition 8.5. Because $t'' < t'$, it follows that an unobserved value for E_i other than its lowest ordered value and the values for $\mathbf{C} \setminus \{E_i\}$ as given by Definition 8.5 could result in a counterfactual explanation.

If we had assumed that E_i is antitone in mode, we can prove the same statement. It follows that a value configuration for $\mathbf{C}$ with $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ can be a counterfactual explanation when subset $\mathbf{S}$ is labeled with a value lower than t'. □

In case the expected value of the target is higher ordered than t a similar proof can be given for the following proposition.

Proposition 8.10. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T\mathbf{e}) = t$ and expected value t', with $t < t'$ and sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, if a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5 is labeled with t'' and $t'' > t'$, a value configuration for $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ is potentially a counterfactual set.

From all previous propositions it follows that we can find all potential counterfactual sets that are inverses of subsets represented in the lattice based on the labels of those subsets. However, we still need to find all counterfactual explanations for those sets. We can do this by looping through all unobserved value configurations for inverses of the subsets with the right label in the lattice and computing the appropriate probabilities. The following proposition states when a value configuration for a subset is counterfactual in the lattice. This statement holds regardless of the ordering of the expected and the most probable value of the target.

Proposition 8.11. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ and expected value t', sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4 and a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, an unobserved value configuration $\mathbf{c}$ for $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ is a counterfactual explanation if all of the following statements are true.

- Based on the label of $\mathbf{S}$, it follows from Propositions 8.8, 8.9 or 8.10 that $\mathbf{C}$ could be a counterfactual set.

- $\top(T|\mathbf{ce}') = t'$, with $\mathbf{e}'$ the observed value configuration for $\mathbf{E} \setminus \mathbf{C}$.
- For all ancestors $\mathbf{A}$ of $\mathbf{S}$ in the lattice for which a counterfactual explanation $\mathbf{c}'$ for $\mathbf{C}' = \mathbf{E} \setminus \mathbf{A}$ is found, $\mathbf{c}$ is not consistent with $\mathbf{c}'$.

Proof. Assume we have a subset in the lattice $\mathbf{S} \subseteq \mathbf{E^L}$ and an unobserved value configuration $\mathbf{c}$ for $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ for which all three statements hold. It follows from the first statement that $\mathbf{C}$ is potentially a counterfactual set.

It follows from the second statement and the definition of the counterfactual explanation that $\mathbf{c}$ is a counterfactual explanation if there is no unobserved value configuration for a subset of $\mathbf{C}$ that is consistent with $\mathbf{c}$ and for which t' is the most probable value given all other evidence as observed.

All subsets of $\mathbf{C}$ are found in the inverses of the ancestors $\mathbf{A}$ of $\mathbf{S}$. It follows from the third statement that the counterfactual explanations that are found for those subsets are not consistent with $\mathbf{c}$. It follows that $\mathbf{c}$ is indeed a counterfactual explanation. □

From the previous propositions we know the subsets with what labels we need to explore to find all potential counterfactual sets and when a value configuration for one of these sets is a counterfactual explanation. We conclude that all counterfactual sets only containing evidence variables represented in the lattice can be found in a lattice by a breadth-first search starting at the top. When a node is found that could potentially provide a counterfactual set, we should first find all value configurations that are counterfactual explanations for this set, before continuing our search to the next node in the lattice. This prevents us from computing the most probable value for a value configuration that is consistent with a counterfactual explanation that is already found. It follows from Proposition 8.11 that if all possible unobserved value configurations for the inverse of a node are counterfactual explanations, none of the children of the node can give a counterfactual explanation, because they would not be minimal. This as a stop criterion when the search is performed. Algorithm 7 dynamically constructs the lattice given by Definition 8.5 while performing a breadth-first search starting at the top of the lattice. In the algorithm there is a loop to check which unobserved value configurations for a subset are a counterfactual explanation. Not all possible values for each subset always need to be checked, because the monotonicity of the evidence can be used to rule out some value configurations if we find that another value configuration is not counterfactual.

Algorithm 7: Computes all counterfactual explanations containing only evidence variables represented in the lattice with a breadth-first search through the lattice for monotone evidence

Input : Target T with most probable value t and expected value t', evidence set $\mathbf{E}$ monotone with T
Output: All counterfactual explanations for sets containing only evidence variables represented in the lattice

```
Counterfactuals = ∅
Q = Queue containing E
while Q not empty do
    S = get first item in Q
    C = E^L \ S
    c = value configuration for C according to Definition 8.5
    e' = the observed value configuration for E \ C
    t'' = ⊤(T|e'c)
    if t'' == t' or (t' > t and t'' > t') or (t' < t and t'' < t') then
        foreach unobserved value configuration c for C do
            if c not consistent with any c' ∈ Counterfactuals and ⊤(T|e'c) == t' then
                Counterfactuals = Counterfactuals ∪ c
            end
        end
    if not all value configurations for C are consistent with counterfactual explanations then
        Put all children of S on Q
    end
end
return Counterfactuals
```

8.4.2 Computing the remaining counterfactual explanations

We proved how we could find all counterfactual explanations based on the variables presented in the lattice, where the excluded evidence variables in $\mathbf{D}$ are not part of. However, we also need to find all counterfactual sets with evidence that was excluded from the lattice.

Based on the label of a node in the lattice, it can be decided if adding a subset $\mathbf{D}' \subseteq \mathbf{D}$ to its inverse is potentially a counterfactual set. We prove this in the following propositions in case the expected value is lower ordered than the most probable value of the target.

Proposition 8.12. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ and expected value t', with $t' < t$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, and $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ for a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, $\mathbf{C} \cup \mathbf{D}'$ with $\mathbf{D}' \subseteq \mathbf{D}$ is potentially a counterfactual set if $\mathbf{S}$ has label t'' with $t'' < t'$.

Proof. Assume we have a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice with its inverse $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ with a label $t'' < t'$. Let $D_i \in \mathbf{D}$ be isotone in mode with the target. The label t'' with $t'' < t'$ for $\mathbf{S}$ was computed for the observed values for $\mathbf{E} \setminus \mathbf{C}$ and an unobserved value configuration for $\mathbf{C}$ as given by Definition 8.5. So the observed value for D_i was used to compute this label. We derive from Definition 8.3 that the observed value is the lowest ordered value for D_i. It now follows from Proposition 7.17 that $\mathbf{C} \cup \{D_i\}$ is potentially a counterfactual set.

If we assume that D_i was antitone in mode with the target, we derive that D_i was excluded from the lattice, because its highest ordered value was observed. It now follows from Proposition 7.20 that $\mathbf{C} \cup \{D_i\}$ is potentially a counterfactual set.

For each individual piece of evidence of $\mathbf{D}$ we have that adding it to $\mathbf{C}$ could create a counterfactual set. It follows that $\mathbf{C} \cup \mathbf{D}'$ could be a counterfactual sets for all subsets $\mathbf{D}' \subseteq \mathbf{D}$. □

If t' would be higher ordered than t, an analogous proof can be given to derive that a subset $\mathbf{D}' \subseteq \mathbf{D}$ can only be added to the inverse of a node with a label higher ordered than the expected value.

Proposition 8.13. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value t', with $t < t'$, sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4, and $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ for a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, $\mathbf{C} \cup \mathbf{D}'$ with $\mathbf{D}' \subseteq \mathbf{D}$ is potentially a counterfactual set if $\mathbf{S}$ has label t'' with $t' < t''$.

We can now decide based on the label of $\mathbf{S}$ if $\mathbf{C} \cup \mathbf{D}'$ is potentially a counterfactual set. For these sets we have to check if they have unobserved value configurations that are counterfactual explanations. In the following statement we prove when an unobserved value configuration for $\mathbf{C} \cup \mathbf{D}'$ is indeed a counterfactual explanation. This proposition holds for all orderings of t and t'.

Proposition 8.14. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$, expected value t', sets $\mathbf{X}$, $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{E^L}$ as given by Definitions 8.1, 8.2, 8.3 and 8.4 and $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ for a subset $\mathbf{S} \subseteq \mathbf{E^L}$ in the lattice given by Definition 8.5, an unobserved value configuration $\mathbf{cd}'$ for $\mathbf{C} \cup \mathbf{D}'$ with $\mathbf{D}' \subseteq \mathbf{D}$ is a counterfactual explanation if all of the following statements hold.

- $\mathbf{C} \cup \mathbf{D}'$ is potentially a counterfactual set according to Proposition 8.12 or 8.13.
- $\top(T|\mathbf{cd}'\mathbf{e}') = t'$, with $\mathbf{e}'$ the observed value configuration for $\mathbf{E} \setminus (\mathbf{C} \cup \mathbf{D}')$.
- For all subsets $\mathbf{C}' \subset \mathbf{C} \cup \mathbf{D}'$ for which a counterfactual explanation $\mathbf{c}'$ is found, $\mathbf{cd}'$ is not consistent with $\mathbf{c}'$.

Proof. Assume we have a subset in the lattice $\mathbf{S} \subseteq \mathbf{E^L}$ and an unobserved value configuration $\mathbf{cd}'$ for $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ and $\mathbf{D}' \subseteq \mathbf{D}$ for which all statements hold.

The computation of the mode in the second statement gives that t' is the most probable value for T given unobserved value configuration $\mathbf{cd}'$ and the observed values for $\mathbf{E} \setminus (\mathbf{C} \cup \mathbf{D}')$. If there is no value configuration consistent with $\mathbf{cd}'$ for a subset of $\mathbf{C} \cup \mathbf{D}'$ that is a counterfactual explanation then $\mathbf{cd}'$ is a counterfactual explanation. This is enforced by the last statement. We conclude that $\mathbf{cd}'$ is minimal and is therefore a counterfactual explanation. □

From the previous propositions we conclude that Algorithm 7 can be extended to also find all counterfactual explanations that do include variables in $\mathbf{D}$. This can be done in the same breadth-first search. If we can derive from the label of the node that adding evidence variables from $\mathbf{D}$ could give a counterfactual set, we should check for all unobserved value configurations if they are counterfactual explanations. This should only be done for value configurations that are not consistent with counterfactual explanations that are already found. This extended algorithm is given in Algorithm 8.

Algorithm 8: Computes all counterfactual explanations a breadth-first search through the lattice for evidence monotone in mode with the target

Input : Target T with most probable value t and expected value t',
evidence set $\mathbf{E}$ monotone with T
Output: All counterfactual explanations

```
Counterfactuals = ∅
Q = Queue containing E
while Q not empty do
    S = get first item in Q
    C = E \ S
    c = value configuration for C according to Definition 8.5
    e' = the observed value configuration for E \ C
    t'' = ⊤(T|e'c)
    if t'' == t' or (t' > t and t'' > t') or (t' < t and t'' < t') then
        foreach unobserved value configuration c for C do
            if c not consistent with any c' ∈ Counterfactuals and ⊤(T|e'c) == t' then
                Counterfactuals = Counterfactuals ∪ c
            else if t' > t and ⊤(T|e'c) > t' or t > t' and ⊤(T|e'c) < t' then
                Find all counterfactual explanations with D' ⊆ D
            end
        end
    if not all value configurations for C are consistent with counterfactual explanations then
        Put all children of S on Q
    end
end
return Counterfactuals
```

8.5 Example

We demonstrate how Algorithm 5 and Algorithm 6 can be combined with Algorithm 8 to find all sufficient and counterfactual sets. We use the BN given in Figure 5 as an example. This network is loosely based on the insurance network as described by Binder et al. (1997).

In this network we have different variables which can be used to determine how probable it is that a driver in a car causes an accident and how severe this accident will be. The target is *Accident* with three possible values *none* (n) < *moderate* (m) < *severe* (s). We have three observable variables that tell us something about the driver; *Age*, *Experience* and *Cautiousness*, with possible values *adolescent* < *adult* < *senior*, *max* 3 *years* < *max* 15 *years* < *more than* 15 *years* and *reckless* < *normal* < *cautious* respectively. These variables are all antitone in mode with the target, it follows that a senior driver who is cautious and has a lot of experience is less likely to cause a severe accident than a younger, reckless driver with less experience. There are also two observable variables about the car; *Mileage* and *Model* with values *max* 5000 *km* < *max* 20000 *km* < *more than* 20000 *km* and *old* < *middle* < *new* respectively. *Mileage* is isotone in mode with the target; with a higher mileage a more severe accident becomes more likely. *Model* is antitone in mode with the target; with a newer model it becomes more likely that the accident is less severe.

We consider a situation with the following evidence. There is an adolescent with 2 years driving experience and with a cautious driving style. He drives in an old model that has driven less than 20000 km. Because he is cautious and his car has a mileage that is not too high, he expects that if he has an accident, that it will be moderate. However, after instantiating the evidence, we find that *severe* is the most probable value for the target. Because all evidence is monotone in mode with the target we can use the algorithms presented in the previous sections to find all explanations.

Example 8.1. Before we run the algorithms we determine for each piece of evidence if it can be excluded from the lattice based on Definitions 8.1, 8.2 and 8.3. *Cautiousness* is antitone in mode and both the observation for *Cautiousness* and the most probable value for the target has the highest ordered value. It follows from Definition 8.3 that *Cautiousness* is excluded from the lattice, because it can not be part of a sufficient set. It seems intuitive that a cautious driver can not be a reason for a severe accident. It follows from Table 3 that in some cases

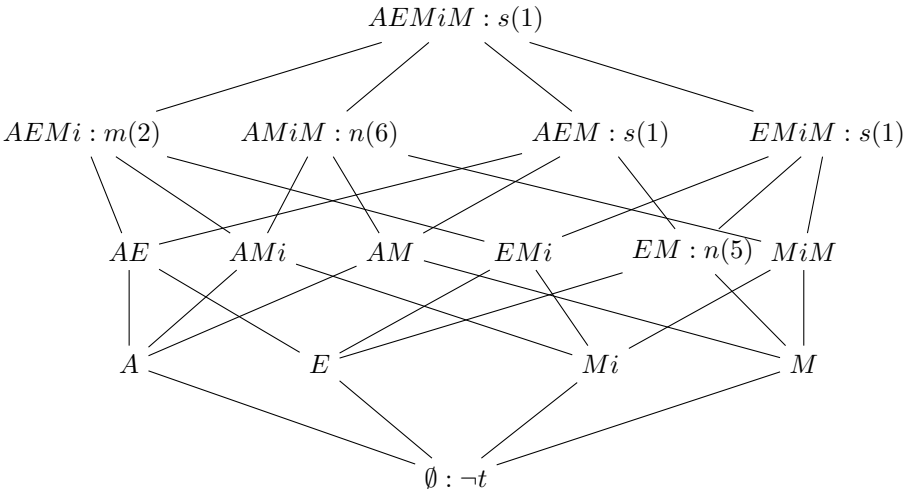


Figure 4: A subset lattice for the Accident network. The labels represent the observations $Age = adolescent$, $Mileage < 20000$, $Model = old$, $Cautiousness = Cautious$, $Experience < 3$. The numbers in the in nodes represent the amount of probabilities that were computed for that node.

Cautiousness can still be part of a counterfactual set. For all other evidence we derive from that it should not be excluded from the lattice. As a result we have $\mathbf{E^L} = \{Age, Experience, Mileage, Model\}, \mathbf{D} = \{Cautiousness\}$ and $\mathbf{X} = \mathbf{R} = \emptyset$.

We now dynamically construct the lattice with $\mathbf{E^L}$. The complete lattice can be found in Figure 4. Only the labels for the subsets that are computed during the search are given. We start the search at the top of the lattice. We find that the top of the lattice is labeled with *severe*. We can not yet make any useful statements about this set, so we put all the children of the top in the queue and continue our search. The first subset in the queue is $\{Age, Experience, Mileage\}$. $\{Model\}$ is the inverse of this subset and it follows from Definition 8.5 that $Model$ should have its highest ordered value when we compute the label for the subset. We find that *moderate* is the most probable value for the target given the observed values for $\{Age, Experience, Mileage\} \cup \{Cautiousness\}$ and the unobserved value *new* for $Model$. *moderate* is the expected value, so we found that $\{Model\}$ is a counterfactual set. We still need to find all unobserved value configurations that are counterfactual explanations. We compute the appropriate mode and conclude that $Model = middle$ and $Model = new$ are both counterfactual explanations. All values for $Model$ are in a counterfactual explanation, it follows that none of the descendants of the current subsets have an inverse that is counterfactual, because it could not be minimal. As a result, we do not put the children of $\{Age, Experience, Mileage\}$ in the queue.

The next subset in the queue is $\{Age, Mileage, Model\}$. $\{Experience\}$ is the inverse of this subset and it follows from Definition 8.5 that it has its highest ordered value when the label for the subset is computed. We find that *none* is the label for this subset. It follows that $Experience > 15$ is not a counterfactual explanation. Because $none < moderate$, it follows from Proposition 8.9 that a lower ordered unobserved value for $Experience$ could be a counterfactual explanation. The most probable value for the target is computed given $Experience < 15$ and the observations for the remaining evidence. We find that *none* is again the most probable value. We find that no value for $Experience$ on its own is a counterfactual explanation. However, it follows from Proposition 8.12 that an unobserved value configuration for $\{Experience, Cautiousness\}$ could be a counterfactual explanation. We need to check all remaining possible value configurations for $Experience$ and $Cautiousness$ to verify this. This gives us four additional probabilities to compute. After checking those probabilities we find that $Experience < 15 \wedge Cautiousness = normal$ is a counterfactual explanation.

The value $Experience > 15$ gives *none* as most probable value for the target given all values of $Cautiousness$. Adding unobserved values for other evidence variables can only give lower ordered values as most probable for the target. So $Experience > 15$ can not be part of a counterfactual explanation and it follows that no descendants of the current subset can give a counterfactual set. So we do not put the children of $\{Age, Mileage, Model\}$ in the queue.

$\{Age, Experience, Model\}$ is the next subset in the queue. We find that *severe* is the label for this subset. However, we are not yet able to conclude whether this subset is sufficient or not, because we do not know if it is minimal. The children of $\{Age, Experience, Model\}$ are put in the queue to verify this.

The next subset is $\{Experience, Mileage, Model\}$. After computing the appropriate probability we find that *severe* is the label for this subset. Again we do not have enough information to conclude that $\{Experience, Mileage, Model\}$ is a sufficient set and we put all its children in the queue.

The following subset in the queue is $\{Age, Experience\}$. For its parent $\{Age, Experience, Mileage\}$ it was concluded that it could not have counterfactual descendants. It can also not be a sufficient set, because one of its parents is labeled with *moderate*. We have that $moderate < severe$ so it follows from Proposition 8.1 that the current subset has label *moderate* or *none*. It follows that the probability for this subset does not have to be computed and we do not put its children in the queue. The same applies to the next set $\{Age, Model\}$.

The next subset in the queue is $\{Experience, Model\}$. The label for this subset is computed given $Age = senior$, $Mileage < 5000$ and the observations for $\{Experience, Model\} \cup \{Cautiousness\}$. We find that *none* is the most probable value for the target. It follows that the value configuration $Age = senior \wedge Mileage < 5000$ is not a counterfactual explanation. Because $none < moderate$, it follows from Proposition 8.9 that another value configuration is potentially a counterfactual explanation. The other unobserved value for $Mileage$ is > 2000. It follows from the monotonicity relation that a counterfactual explanation can not contain $Mileage > 2000$. Age has value *Adult* that still could provide a counterfactual explanation in combination with $Mileage < 500$. We find that *none* is the most probable value given $Age = Adult \wedge Mileage < 5000$ and the observations for the remaining evidence. We conclude that $\{Age, Mileage\}$ is not a counterfactual set. However, we derive from Proposition 8.12 that $\{Age, Mileage, Cautiousness\}$ is potentially a counterfactual set. We need to compute additional probabilities to determine what unobserved value configurations for this set are counterfactual explanations. First we find that *none* is the most probable value given $Mileage < 500 \wedge Age = adult \wedge Cautiousness = normal$ and the observed values for the remaining evidence. So this value configuration is not a counterfactual explanation. Because Age is antitone in mode with the target, we can derive that $Mileage < 500 \wedge Age = senior \wedge Cautiousness = normal$ is not a counterfactual explanation. Secondly we compute that $Mileage < 500 \wedge Age = adult \wedge Cautiousness = reckless$ has *moderate* as most probable value given the observations for the remaining evidence. We conclude that this unobserved value configuration is a counterfactual explanation. At last we compute the appropriate mode for $Mileage < 500 \wedge Age = senior \wedge Cautiousness = normal$ and find that it is not a counterfactual explanation. We had to compute three additional probabilities to find the counterfactual explanations with $Cautiousness$.

none is the most probable value for all value configurations for $\{Age, Mileage\}$. The unobserved values for evidence in the lattice will can only make *none* more likely, so it follows that no superset of $\{Age, Mileage\}$ can be a counterfactual set. As a result, we do not put the children of $\{Experience, Model\}$ in the queue.

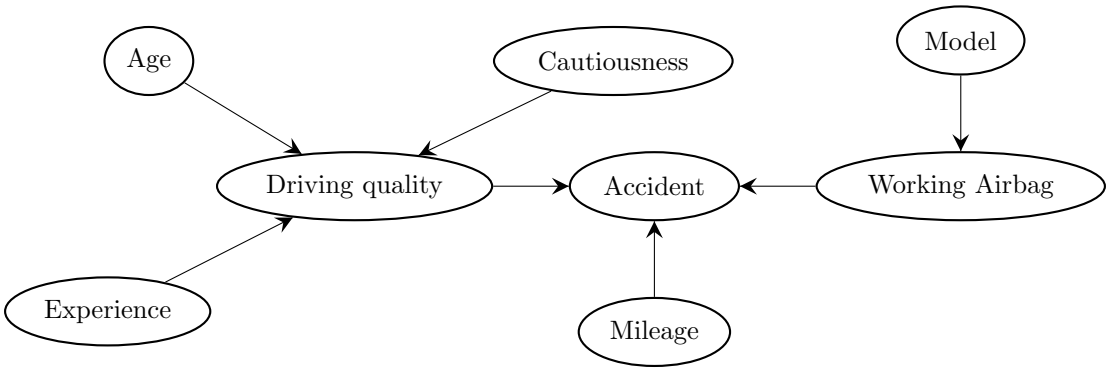
The next subset in the queue is $\{Mileage, Model\}$. For one of the parents of this subset it was concluded that no descendants were counterfactuals, so no probabilities are computed for this subset. The queue is now empty and the algorithm terminates.

We have found the following sufficient explanations; $Age = adolescent \wedge Experience < 3 \wedge Model = old$ and $Mileage < 2000 \wedge Experience < 3 \wedge Model = old$. These are the nodes in the lattice that are labeled with s and do not have any children with this label. The counterfactual explanations we found are $Model = normal$ and $Model = new$, $Experience < 7 \wedge Cautiousness = normal$ and $Mileage < 500 \wedge Age = adult \wedge Cautiousness = reckless$. 16 probabilities were computed to find these explanations.

All the sufficient explanations should be combined with all the counterfactual explanations to construct all contrastive, counterfactual explanations. In Section 9 we see how these explanations are presented to the user.

8.6 Observations

We saw in the previous sections that all counterfactual sets can be found with one breadth-first search through the lattice. To find all sufficient sets some additional checks are done after the breadth-first search. In finding the sufficient sets there is a difference in complexity when the most probable value of the target is its highest or lowest ordered value and when it it not the highest or lowest ordered value. In the first situation, one mode needs to be computed for every node in the lattice and all sets can be found with one breadth-first search. This gives us 2^n modes to compute in the worst-case scenario, with n the size of the set with evidence variables represented in the lattice. For the same reasoning as given in case of binary valued evidence, it is unlikely that the modes for all subsets in the lattice needs to be computed when searching for the sufficient sets. If the most probable value is not the highest or lowest ordered value for the target, one additional mode needs to be computed for every node in the lattice and some additional computations are necessary after the breadth-first search. To find all sufficient sets with only evidence represented in the lattice, a maximum of 2^{n+1} modes need to be computed in case the most probable value for the target is not its highest or lowest ordered value. Again this worst-case scenario is not likely to occur often.



$P(A = Adolescent) = 0.2$ $P(A = Adult) = 0.6$ $P(A = Senior) = 0.2$

$P(C = Reckless) = 0.4$ $P(C = Normal) = 0.5$ $P(C = Cautious) = 0.1$

$P(E < 3) = 0.5$ $P(E < 7) = 0.45$ $P(E > 7) = 0.05$

$P(Mi < 5000) = 0.1$ $P(Mi < 20000) = 0.5$ $P(Mi > 20000) = 0.4$

$P(M = Old) = 0.75$ $P(M = Normal) = 0.2$ $P(M = New) = 0.05$

$P(Wa = true|M = Old) = 0.1$ $P(Wa = false|M = Old) = 0.9$
$P(Wa = true|M = Normal) = 0.8$ $P(Wa = false|M = Normal) = 0.2$
$P(Wa = true|M = New) = 0.95$ $P(Wa = false|M = New) = 0.05$

Figure 5: A BN representing the probability that a driver has a car accident, based on variables about the driver and his car. The probabilities for the *Accident* and *Driving Quality* can be found in Table 4

Experience	<3								
Cautiousness	*Reckless*			*Normal*			*Cautious*		
Age	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*
$P(DQ = poor)$	0.9	0.85	0.6	0.85	0.8	0.55	0.8	0.7	0.5
$P(DQ = normal)$	0.09	0.13	0.38	0.1	0.13	0.35	0.15	0.2	0.35
$P(DQ = excellent)$	0.01	0.02	0.02	0.05	0.07	0.1	0.5	0.1	0.15

Experience	<15								
Cautiousness	*Reckless*			*Normal*			*Cautious*		
Age	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*
$P(DQ = poor)$	0.9	0.85	0.6	0.85	0.8	0.55	0.8	0.7	0.5
$P(DQ = normal)$	0.09	0.13	0.38	0.1	0.13	0.35	0.15	0.2	0.35
$P(DQ = excellent)$	0.01	0.02	0.02	0.05	0.07	0.1	0.5	0.1	0.15

Experience	>15								
Cautiousness	*Reckless*			*Normal*			*Cautious*		
Age	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*	*Ado*	*Adult*	*Sen*
$P(DQ = poor)$	0.1	0.05	0.05	0.1	0.08	0.05	0.05	0.03	0.01
$P(DQ = normal)$	0.6	0.3	0.2	0.4	0.22	0.2	0.35	0.22	0.19
$P(DQ = excellent)$	0.3	0.65	0.75	0.5	0.7	0.75	0.6	0.75	0.8

WorkingAirbag	*False*								
Mileage	<5000			<20000			>20000		
DrivingQuality	*P*	*N*	*E*	*P*	*N*	*E*	*P*	*N*	*E*
$P(Ac = none)$	0.27	0.57	0.57	0.15	0.4	0.45	0.13	0.3	0.5
$P(Ac = moderate)$	0.38	0.28	0.3	0.35	0.33	0.35	0.32	0.4	0.4
$P(Ac = severe)$	0.35	0.15	0.13	0.5	0.27	0.2	0.55	0.3	0.1

WorkingAirbag	*True*								
Mileage	<5000			<20000			>20000		
DrivingQuality	*P*	*N*	*E*	*P*	*N*	*E*	*P*	*N*	*E*
$P(Ac = none)$	0.52	0.7	0.85	0.25	0.55	0.65	0.2	0.5	0.54
$P(Ac = moderate)$	0.18	0.25	0.14	0.05	0.3	0.25	0.5	0.38	0.36
$P(Ac = severe)$	0.3	0.05	0.01	0.25	0.15	0.1	0.3	0.12	0.1

Table 4: The probabilities for *Accident* and *Driving Quality* in the BN given in Figure 5.

When finding all counterfactual sets one mode is computed to decide if a set could be counterfactual. In the worst-case scenario 2^n modes need to be computed to decide for all sets if they could be counterfactual or not, where n is the size of the set with evidence variables represented in the lattice. When a potential counterfactual sets is found, it needs to be computed what unobserved value configurations are counterfactual explanations. This is where the real complexity lies. Given a counterfactual set of size $m \leq n$ with k probable values for all evidence in the set, there are $m^{(k-2)}$ unobserved value configurations for which the most probable value of the target is not yet computed. However, it is not always necessary to compute all modes for all these value configurations, because the monotonicity relation between the evidence and the target can be used to derive if a value configuration is a counterfactual explanation.

We also provided some rules that could be used to exclude certain evidence from the lattice. If it could be derived for l evidence variables that they could be excluded from the lattice, the lattice would then have $2^{(n-l)}$ nodes instead of 2^n, with n the size of the complete evidence set. This could significantly decrease the number of modes to compute. However, some excluded evidence variables could still be part of the sufficient or counterfactual set of the explanation. From several propositions proposed in the previous sections it follows that variables can only be added to subsets in the lattice which are labeled with specific values for the target. In those cases additional modes need to be computed to make sure all explanations are found. However, those modes would also have been computed when the evidence variables would not have been excluded from the lattice. In the worst-case scenario all nodes in the lattice have a label for which it is necessary to compute the additional probabilities for the excluded evidence variables. In this situation finding all explanations in a lattice where variables are excluded has the same complexity as finding all explanations in a lattice where the variables were not excluded. In all other situations less computations needs to be done in a lattice with excluded evidence variables.

9 Presenting the explanation to the user

In the previous sections we saw how all sufficient and counterfactual explanations are computed in a lattice in case all evidence variables are binary valued or all evidence is monotone in mode with the target. All sufficient explanations can now be combined with all counterfactual explanations to construct all contrastive, counterfactual explanations. In the example in Section 8.5 we found 2 sufficient explanations and 4 counterfactual explanations, which gives a total of 8 contrastive, counterfactual explanations. If we have a BN with more evidence, we expect to find even more possible explanations. We do not think it is useful to present all these explanations to the user. In Section 9.1 we describe two methods of how a selection can be made of all explanations to present to the user. One method is used when the user has a preference for explanations that are easily understandable, the other is used when the user has a preference for explanations that are more probable.

After a selection of the explanations is made, the question raises of how these explanations should be presented to the user. In Section 9.2 we give different templates of how the explanations can be given to the user in a textual way. In Section 9.3 we give two examples of how the explanations can be further adjusted to the wants and needs of the user. We demonstrate in Section 9.3 how the selection of the explanations and the additional information is presented to the user in practice.

9.1 Selecting explanations

When all explanations are found in the lattice, we do not want to present them all to the user, but we want to make a selection of explanations that are most useful for the user. We define two different ways in which a selection of the explanations can be made. First of all, a selection can be made based on how easily it is to understand an explanation. For example, it takes less effort to understand an explanation with less variables than an explanation with more variables. Secondly, a selection can be made based on how convincing an explanation is. We say that an explanation is more convincing if it is more probable. Based on the preference of the user, we order all explanations by how understandable or convincing they are. The top 3 most understandable or convincing explanations will then be given to the user. If the user wants more explanations, the following 3 are given to him, etcetera.

In Section 9.1.1 we define what an understandable explanation is. We also explain how the problem of finding all explanations can be simplified if sufficient and counterfactual sets are not explored that are expected to be more difficult to understand. In Section 9.1.2 we define when an explanation is convincing. To illustrate the definitions of understandable and convincing explanations, we revisit the swimming pool example given in Section 2.2. The following case will be used throughout both sections to clarify the definitions.

Contrastive, counterfactual explanations	
Sufficient explanation	**Counterfactual explanation**
$Temp < 25 \wedge Sunny = false$	$Temp > 30$
$Temp < 25 \wedge Friend = false$	$Temp > 30$
$Temp < 25 \wedge Sunny = false$	$Temp = 25 - 30 \wedge Sunny = true$
$Temp < 25 \wedge Friend = false$	$Temp = 25 - 30 \wedge Friend = true$
$Temp < 25 \wedge Sunny = false$	$Temp = 25 - 30 \wedge Friend = true$
$Temp < 25 \wedge Friend = false$	$Temp = 25 - 30 \wedge Sunny = true$
$Temp < 25 \wedge Sunny = false$	$Sunny = true \wedge Friend = true$
$Temp < 25 \wedge Friend = false$	$Sunny = true \wedge Friend = true$

Table 5: All contrastive, counterfactual explanations for the case where we did not go to the pool based on the following observations *Temperature* < 25, *Sunny* $=$ *false* and *Friend* $=$ *false*, ordered by their understandability.

Example 9.1. Consider a day where it is colder than 25 degrees, it is not sunny and our friend does not go to the pool. Based on this observations, we decide not to go to the pool. We observe that there are two sufficient explanations for our current decision; $Temp < 25 \wedge Sunny = false$ and $Temp < 25 \wedge Friend = false$.

We also observe there are several counterfactual explanations for why we did not go to the pool, these are; $Sunny = true \wedge Friend = true$, $Temp = 25 - 30 \wedge Sunny = true$, $Temp = 25 - 30 \wedge Friend = true$ and $Temp > 30$. These counterfactual explanations can be combined with all sufficient explanations to provide contrastive, counterfactual explanations. All these explanations can be found in Table 5.

9.1.1 Understandable explanations

We define an understandable explanation, as an explanation that is simple and does not contain too much information, so the cognitive load to understand it is not too high. People generally prefer simple explanations, where fewer causes are given for a result (Miller, 2019). Based on these observations we define two different properties, which we can use to determine how understandable an explanation is. These properties are; *conciseness* and *uniformity*.

Conciseness relates to the total number of variables in the sufficient and the counterfactual set of the explanation. A simple explanation with less variables is more concise and thus preferred over a complex explanation with more variables. This leads to the following definition of conciseness;

Definition 9.1. (Conciseness) Given a contrastive, counterfactual explanation according to Definition 4.4, the conciseness of the explanation is the size of $\mathbf{S} \cup \mathbf{C}$, where an explanation is considered more concise if this size is smaller.

Of the explanations given in Table 5 the explanations in the first block are the most concise, because they only have three variables in both sets together, where the other explanations have four.

The uniformity-property is related to conciseness because it also looks at the variables in both sets. However, it looks at the number of overlapping variables between the sufficient and counterfactual set. An explanation with a greater overlap in variables between the sets is preferred over an explanation with less overlap. We consider an explanation where the sets are more overlapping to be more uniform and therefore better interpretable.

Definition 9.2. (Uniformity) Given a contrastive, counterfactual explanation according to Definition 4.4, the uniformity of the explanation is the size of $\mathbf{S} \cap \mathbf{C}$, where an explanation is considered more uniform if this size is higher.

If we look at the explanations in Table 5 we see that the explanations in the second block are more uniform than those in the third block. The explanations in the second block have two overlapping variables between the counterfactual set and the sufficient set, where the explanations in the third block only have one overlapping variable. We think that the explanations in the second block are therefore more uniform and better understandable than the explanations in the last block.

If we want to order explanations based on their understandability, we order them first by conciseness, then by uniformity. We order them first by conciseness, because we want a smaller number of variables in the explanation. If

we would order them in the other way, explanations with more variables in both sets are preferred over explanations with less variables, because explanations with less variables automatically have a lower uniformity.

If a user prefers understandable explanations over convincing explanations, we could simply compute all explanations from the lattice, then order them as described and only present the most understandable explanations to the user. However, when the user prefers understandable explanations we could also simplify the algorithm that finds all explanations. We propose two different ways to do this. The first method simplifies the algorithm by earlier stopping the search for sufficient or counterfactual sets that contain evidence that is not represented in the lattice. We observe that sufficient and counterfactual sets that contain evidence that is not represented in the lattice place a lower bound on the conciseness and an upper bound on the uniformity for all contrastive, counterfactual explanations they are part of.

Proposition 9.1. Given a subset $\mathbf{S}$ in the lattice given by Definition 8.5 and a set $\mathbf{R}$ as given by Definition 8.2, a contrastive, counterfactual explanation with a sufficient set $\mathbf{S} \cup \mathbf{R}'$ with $\mathbf{R}' \subseteq \mathbf{R}$ has a conciseness of at least $|\mathbf{R}'|$ and a uniformity of at most $|\mathbf{S}|$.

Proof. Given is a contrastive, counterfactual explanation with a sufficient set $\mathbf{S} \cup \mathbf{R}'$ and a counterfactual set $\mathbf{C}$.

For each evidence variable in $\mathbf{R}'$ it was derived that it could not be part of a counterfactual set. So we have $\mathbf{R}' \cap \mathbf{C} = \emptyset$. It now follows that $|\mathbf{S} \cup \mathbf{R}' \cup \mathbf{C}| \geq |\mathbf{R}'|$. It follows that the conciseness of all explanations with $\mathbf{S} \cup \mathbf{R}'$ as sufficient set is higher than the size of $\mathbf{R}'$.

Because $\mathbf{R}' \cap \mathbf{C} = \emptyset$, it also follows that $|(\mathbf{S} \cup \mathbf{R}') \cap \mathbf{C}| \leq |\mathbf{S}|$. We conclude that the uniformity of all explanations with $\mathbf{S} \cup \mathbf{R}'$ is at most $|\mathbf{S}|$. □

The same is true for a counterfactual set that contains evidence variables that were excluded from the lattice.

Proposition 9.2. Given evidence $\mathbf{E}$, a subset $\mathbf{S}'$ in the lattice given by Definition 8.5 with $\mathbf{C} = \mathbf{E}^{\mathbf{L}} \setminus \mathbf{S}'$ and a set $\mathbf{D}$ as given by Definition 8.3, a contrastive, counterfactual explanation with a counterfactual set $\mathbf{C} \cup \mathbf{D}'$ with $\mathbf{D}' \subseteq \mathbf{D}$ has a conciseness of at least $|\mathbf{D}'|$ and a uniformity of at most $|\mathbf{C}|$.

Proof. Given is a contrastive, counterfactual explanation with a sufficient set $\mathbf{S}$ and a counterfactual set $\mathbf{C} \cup \mathbf{D}'$.

For each evidence variable in $\mathbf{D}'$ it was derived that it could not be part of a sufficient set. So we have $\mathbf{D}' \cap \mathbf{S} = \emptyset$. It now follows that $|\mathbf{S} \cup \mathbf{C} \cup \mathbf{D}'| \geq |\mathbf{D}'|$. It follows that the conciseness of all explanations with $\mathbf{C} \cup \mathbf{D}'$ as counterfactual set is higher than the size of $\mathbf{D}'$.

Because $\mathbf{D}' \cap \mathbf{S} = \emptyset$, it also follows that $|\mathbf{S} \cap (\mathbf{C} \cup \mathbf{D}')| \leq |\mathbf{C}|$. We conclude that the uniformity of all explanations with $\mathbf{C} \cup \mathbf{D}'$ is at most $|\mathbf{C}|$. □

These propositions can be used to deduce if a better understandable explanation than those already found can be created with the evidence variables excluded from the lattice. If continuing the search will only result in explanations less understandable than those already found, the search can be terminated.

The second method of reducing the complexity is by earlier terminating the search for counterfactual explanations with only evidence variables represented in the lattice. If we find that the inverse of a subset is a counterfactual set, we do not have to explore the children of this subset, because those children can only provide explanations that are less concise. We prove this in the following proposition.

Proposition 9.3. Given evidence $\mathbf{E}$ that is monotone in mode with target T, $\top(T|\mathbf{e}) = t$ and expected value t', an unobserved value configuration for the inverse of a descendant of a set $\mathbf{S} \subseteq \mathbf{E}^{\mathbf{L}}$ in the lattice given by Definition 8.5 can not be part of an explanation that is less concise than an explanation with a counterfactual explanation found for $\mathbf{C} = \mathbf{E}^{\mathbf{L}} \setminus \mathbf{S}$.

Proof. Given is a value configuration $\mathbf{c}$ for $\mathbf{C} = \mathbf{E}^{\mathbf{L}} \setminus \mathbf{S}$ that is a counterfactual explanation, with $\mathbf{S}$ a set in the lattice given by Definition 8.5. If $\mathbf{c}$ is combined with a sufficient explanation to create a contrastive, counterfactual explanation, it has a conciseness of at least $|\mathbf{C}|$.

Given a descendant of $\mathbf{S}$ with its inverse $\mathbf{C}'$, we have $\mathbf{C} \subset \mathbf{C}'$. It follows that $|\mathbf{C}| < |\mathbf{C}'|$. It follows that combining $\mathbf{C}'$ with a sufficient set always gives a conciseness that is at least as high as the conciseness of $\mathbf{C}$ and that same sufficient set. □

If we find a counterfactual set for a node in the lattice, we can use this proposition to decide to not explore the children of the node. The explanations that are found after the described methods are applied to the algorithm are ordered based on their understandability. The top most of those explanations are then presented to the user.

9.1.2 Convincing explanations

We consider explanations that are more probable, more convincing for the user. If an explanation is given to the user that is highly improbable, the user will be less convinced by the explanation and as a result can be less trusting of the system. We define three different probabilities on which we base how convincing an explanation is; the *sufficient conditional probability*, *counterfactual probability* and the *counterfactual conditional probability*. We consider an explanation where these probabilities are all high to be more convincing. If the user has a preference for convincing instead of understandable explanations, we order them based on these probabilities and only give the most probable ones. The definitions of the probabilities and how to order them is explained below.

We want to provide the user with the sufficient sets that have the strongest influence on the most probable value of the target. A sufficient set that has a strong influence on the target will provide a more convincing explanation to the user than a sufficient set with a weaker influence. We do not take the probability of actually observing the sufficient set, $P(\mathbf{s})$, into account. Because we already observed the sufficient set, $P(\mathbf{s})$ will not give any measure of the influence of this set on the target. This leads to the following definition for the sufficient conditional probability;

Definition 9.3. (Sufficient conditional probability) Given a contrastive, counterfactual explanation according to Definition 4.4 with a sufficient set $\mathbf{S} \subseteq \mathbf{E}$ for evidence $\mathbf{E}$ and target T with $\top(T\mathbf{e}) = t$, the sufficient probability of the explanation is given by $P(t|\mathbf{s})$.

Given Example 9.1, we have two sufficient sets to explain $Pool = false$, the probabilities of these sets are; $P(Pool = false|Temp < 25,\ Sunny = false) = 0.68$ and $P(Pool = false|Temp < 25,\ Friends = false) = 0.64$. We notice that observing the values for the sufficient set $\{Temp,\ Sunny\}$ has the strongest influence on the target, so this gives the most convincing explanation.

For the counterfactual explanation we take two different probabilities into account. First of all we prefer a counterfactual explanation where the probability of observing the value configuration for this explanation is high. Providing the user with a counterfactual explanation that is more probable to occur is a more convincing explanation for the user than a counterfactual explanation that is less likely to be observed. However, we do not want to look at the probability $P(\mathbf{c})$ for a counterfactual set $\mathbf{C}$, but we will consider the probability of observing $\mathbf{c}$ while the other evidence is still present. In this way a smaller counterfactual set has no advantage over a larger set. We do not necessarily want to prefer smaller counterfactual sets, but we want to prefer explanations where the assigned values for the variables were the most likely to be observed in addition to the other evidence.

This leads us to the following definition for the counterfactual probability;

Definition 9.4. (Counterfactual probability) Given a contrastive, counterfactual explanation according to Definition 4.4 with a counterfactual set $\mathbf{C} \subseteq \mathbf{E}$ for evidence $\mathbf{E}$ and target T with $\top(T\mathbf{e}) = t$ and expected value t', $t' \neq t$, the counterfactual probability of the explanation is given by $P(\mathbf{ce}')$ with $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

Given Example 9.1, the counterfactual explanation $Sunny = true \wedge Friend = true$ is the most likely to be observed with $P(Temp < 20,\ Sunny = true,\ Friend = true) = 0.12$. So this explanation has a preference over the others, according to the counterfactual probability.

Secondly, we prefer a counterfactual explanation that has a strong influence on the expected target value t'. An explanation where t' is more probable, will probably be more convincing to the user. Again, we want to take the current situation into account, so we look at the probability of t' given the value configuration of the counterfactual explanation and the other evidence.

Definition 9.5. (Counterfactual conditional probability) Given a contrastive, counterfactual explanation according to Definition 4.4 with a counterfactual set $\mathbf{C} \subseteq \mathbf{E}$ for evidence $\mathbf{E}$ and target T with $\top(T\mathbf{e}) = t$ and expected value t', $t' \neq t$, the counterfactual conditional probability of the explanation is given with $P(t'|\mathbf{ce}')$ with $\mathbf{e}'$ the observed value configuration for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$.

Given Example 9.1, the counterfactual explanation $Temp > 30$ is preferred over the others because the counterfactual conditional probability, $P(Pool = true|Temp > 30,\ Sunny = false,\ Friend = false)$, is the highest given this counterfactual explanation.

All counterfactual, contrastive explanations with the appropriate probabilities for the swimming pool example are given in Figure 6. Now that we have three probabilities to determine how convincing an explanation is, we need a way to combine those probabilities into one term to be able to order them and select the most probable. This term we call the probability rating of the explanation. Because we deem all probabilities as important as the others and we want a simple way of combining them, we compute the probability rating by multiplying the three

Contrastive, counterfactual explanations					
Sufficient explanation	$P(t\|\mathbf{s})$	**Counterfactual explanation**	$P(\mathbf{c}, \mathbf{e}')$	$P(t'\|\mathbf{c}, \mathbf{e}')$	R
$Temp < 25 \wedge Sunny = false$	0.68	$Sunny = true \wedge Friend = true$	0.120	0.70	0.0571
$Temp < 25 \wedge Sunny = false$	0.68	$Temp = 25 - 30 \wedge Sunny = true$	0.108	0.75	0.0551
$Temp < 25 \wedge Friend = false$	0.64	$Sunny = true \wedge Friend = true$	0.120	0.70	0.0538
$Temp < 25 \wedge Friend = false$	0.64	$Temp = 25 - 30 \wedge Sunny = true$	0.108	0.75	0.0518
$Temp < 25 \wedge Sunny = false$	0.68	$Temp > 30$	0.048	0.77	0.0251
$Temp < 25 \wedge Friend = false$	0.64	$Temp > 30$	0.048	0.77	0.0237
$Temp < 25 \wedge Sunny = false$	0.68	$Temp = 25 - 30 \wedge Friend = true$	0.048	0.70	0.0228
$Temp < 25 \wedge Friend = false$	0.64	$Temp = 25 - 30 \wedge Friend = true$	0.048	0.70	0.0215

Table 6: All contrastive, counterfactual explanations for the case where we did not go to the pool based on the following observations *Temp < 25*, *Sunny = false* and *Friend = false*, with their probabilities and ordered by probability rating.

given probabilities. The explanation with the highest rating is the most convincing, because it has the highest combined probabilities.

When all explanations are computed in the lattice, the counterfactual conditional probability is computed. However, an additional probability needs to be computed for each sufficient and counterfactual explanation that is found in the lattice. After these are computed, the explanations are ordered based on how convincing they are. The most convincing explanations are then given to the user.

If the user does not necessarily care for the most convincing explanations, but would like to have an indication of the probability of the sufficient set, we could use an adjusted probability instead of the sufficient conditional probability. In this way no additional probabilities have to be computed for each sufficient explanation. Given evidence that is monotone in mode with the target, for each subset $\mathbf{S}$ in the lattice the most probable value is computed for the target given the observations for a set $\mathbf{S}$ and the values for $\mathbf{E} \setminus \mathbf{S}$ as given by Definition 8.5. If it turns out that $\mathbf{S}$ is a sufficient set, this probability can be used instead of the sufficient conditional probability as an indication of the influence of $\mathbf{S}$ on the target. This reduces the number of probabilities that need to be computed for each explanation, but still gives an indication of how probable the explanation is.

9.2 Presenting the explanation

After we have selected what explanations to present to the user, we should decide how we present these explanations to the user. We give two ways of how this can be done.

First of all, the explanations can be given to the user in a schematic way. In case the user prefers understandable explanations a table similar to Table 5 is given to the user. If he prefers convincing explanations a table similar to Table 6 is given, where the appropriate probabilities are presented along with the sufficient and counterfactual explanations. These tables should only contain a selection of all explanations. In case the user wants to see more explanations, the tables could be extended.

This way of presenting the explanations, gives the user a quick overview of all sufficient and counterfactual sets. However, the user should understand the definition of both the sufficient and counterfactual explanation to be able to derive the information presented in the tables. It would be easier for the user to understand the explanations if they were given in a textual way.

We present different templates that can be used to give the sufficient or counterfactual explanation in a textual way. With the templates for the sufficient set we want to convey three different things to the user. Firstly, we want to give the most probable value for the target. Secondly, we want to give the observations for the sufficient set. At last we want to communicate the definition of a sufficient set as clearly as possible to the user in an informal way. With these requirements in mind we constructed the following template.

Template 9.1. Given a target T with most probable value t, a sufficient explanation $E_1 = e_1 \wedge ... \wedge E_n = e_n$ can be given to the user with the following template:

If only $E_1 = e_1, ..., E_n = e_n$ was observed t would always be the most probable value for T regardless of the values for the other evidence.

In case the user wanted a convincing explanation, the template can be extended to include the sufficient conditional probability.

Template 9.2. Given a target T with most probable value t, a sufficient explanation $\mathbf{s}: E_1 = e_1 \wedge ... \wedge E_n = e_n$ can be given to the user with the following template:

If only $E_1 = e_1, ..., E_n = e_n$ was observed t would have a probability of $P(t|\mathbf{s})$. This would always be the most probable value for T regardless of the values for the other evidence.

The templates for the counterfactual explanation should convey similar information as the templates for the sufficient explanations. They should at least contain the expected value, the evidence with the unobserved value configuration and an informal definition of the counterfactual set. This resulted in the following template.

Template 9.3. Given a target T with expected value t', a counterfactual explanation $E_1 = e_1 \wedge ... \wedge E_n = e_n$ can be given to the user with the following template:

If $E_1 = e_1, ..., E_n = e_n$ was observed instead of the actual values and the values for the other variables would stay the same, t' would be the most probable value for T.

Again, different templates should be used when the user wants a convincing explanation.

Template 9.4. Given evidence $\mathbf{E}$ and target T with expected value t', a counterfactual explanation $c: E_1 = e_1 \wedge ... \wedge E_n = e_n$ and $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$ can be given to the user with the following template:

The probability of observing $E_1 = e_1, ..., E_n = e_n$ is $P(\mathbf{ce}')$ given the values for the other variables. If this was observed t' would be the most probable value for T with a probability of $P(t'|\mathbf{ce}')$.

The whole contrastive, counterfactual explanation is given to the user by first filling in the template for the sufficient explanation and then the template for the counterfactual explanation. Depending on the domain in which the explanation is given and the meaning of the variables in the BN, the wording of these templates can be adjusted.

9.3 Adjusting the explanation

In this section two different examples of how the explanations can be further adjusted to the needs and preferences of the user. In Section 9.3.1 we explain how a selection of the evidence can be used to compute all explanation. This could be useful if the user wants to change his behaviour based on the results of the BN. In Section 9.3.2 we show how additional information can be provided about certain variables based on the expectations of the user.

9.3.1 Selecting evidence

In some domains, a specific target value is desired over the other values for the target. Consider for example the Car crash BN given in Figure 5 and say *Accident* is the target. A car driver can use this BN to determine how likely he is to get in a car accident and how severe this accident will be, based on variables about the driver and the car. Obviously the driver prefers that he is not likely to be in a car crash. If the BN indicates otherwise, he could use the counterfactual explanation to determine how he can change his behaviour to make an accident less likely. However, not all variables are as easily adjustable as others. For example, the variable *Cautiousness* can easily be changed by the driver if he adjusts his driving style. On the other hand, the driver can not take action to change the value of *Age*.

If a user expresses that he wants to use the explanation to change his behaviour, the user should indicate which variables can be adjusted by him. Say the user indicates that the evidence $\mathbf{E}' \subset \mathbf{E}$ is changeable. Given a set $\mathbf{E}'$ of changeable evidence, we then define a counterfactual explanation as an unobserved value configuration for $\mathbf{C} \subseteq \mathbf{E}'$ for which all conditions in Definition 4.1 hold.

The search for all sufficient sets now works the same as before. When the counterfactual sets are searched in the lattice, only the subsets $\mathbf{S}$ in the lattice needs to be explored where $\mathbf{C} = \mathbf{E^L} \setminus \mathbf{S}$ only contains evidence in $\mathbf{E}'$. For those subsets it is determined in the same way as given before if they are counterfactual. The value configurations that are counterfactual explanations are also found in the same way.

There is not always a counterfactual explanation when the adjusted definition of counterfactual sets is used, because the set $\mathbf{E}'$ of changeable evidence provided by the user could be too small. If this is the case, the user should reconsider what evidence to put in $\mathbf{E}'$, after which a new search can be performed.

9.3.2 Verifying the results

The contrastive, counterfactual explanation can be used when the expected target value of the user does not align with the most probable target value indicated by the BN. In this case we want to ask the user on what observations his expectations were based. He can then indicate what observations were decisive for his expectation. Based on this selection we provide the user with some additional information about why his expectations were not met. In this way the user learns more about how the different variables are represented in the BN, which could gain more trust by the user. Alternatively, it could be the case that the explanation and the additional information points out some relations or indepencies in the BN that are not represented correctly. This could help to further improve the BN.

Say the user indicates that the observation for evidence E_i was why he expected value t' for the target. Based on how E_i is used by the algorithm we provide the user with additional information. The following conclusions can be derived after all explanations are computed by looking at if and how E_i is used in the explanation.

First of all, it could be decided before the algorithm is run that $E_i \in \mathbf{X}$, $E_i \in \mathbf{R}$ or $E_i \in \mathbf{D}$. If $E_i \in \mathbf{X}$ or $E_i \in \mathbf{R}$, it follows that other values for E_i than observed make the most probable value for the target more likely. If $E_i \in \mathbf{X}$ or $E_i \in \mathbf{D}$, it follows that other values for E_i than observed can not make the expected value more probable. In both cases it follows that the user probably has the right expectations of how E_i interacts with the target. However, the influence of E_i is suppressed by the influence of the other evidence on the target. This conclusion should be communicated to the user. This can for example be done in the following way: *You probable have the right expectations of how E_i influences the target. However, this influence was suppressed by the other observations.*

In case E_i is not excluded from the lattice, there are multiple different options of how it is used in the explanations. E_i is either not in any sufficient or counterfactual explanation or it is included in one or both of them. Based on its inclusion in one of the sets, we could derive the following information. First of all, consider the situation where E_i is not part of any sufficient and not of any counterfactual set. That E_i is not in a sufficient set indicates two things. The current observation for E_i is not a reason for the most probable value for the target, because E_i makes this value less probable or the current observation for E_i makes the most probable value for the target more likely, but has only a minimal influence. That E_i is not included in a counterfactual sets indicates that changing the value for E_i does not make the expected value more probable or only by a small measure. It follows that the user does not necessarily have wrong expectations of how E_i and the target interact, however the observation was not as important as the user thought. This can be communicated to the user by stating; *E_i only has a small or a negative influence on the probabilities of both t and t', but your expectations of how E_i and T interact are not necessarily wrong.*

Secondly, it can be the case that E_i is part of a sufficient explanation but not of a counterfactual explanation. Because E_i is part of a sufficient explanation, it follows that the current observation for E_i has a positive influence on the most probable value for T. However, because E_i is not included in any counterfactual explanation it follows that changing its value does not necessarily make the expected value more likely or only has a small influence. It follows that the user has a wrong expectation of how the observed value for E_i influences the target. However, he was right to believe that other values for E_i do not make his expectations more likely. This conclusion can be given to the user by stating *Another observation for E_i will only have a minimal or negative influence on the probability of t'. However, the current observation was important for the probability of t.*

Now consider the situation where E_i is not part of a sufficient explanation but is part of a counterfactual explanation. Because E_i is not part of a sufficient explanation, we conclude that E_i has a negative or a small positive influence on the most probable value. That E_i is part of a counterfactual set indicates that another value for E_i makes the expected value for the target more likely. It follows that the user was right to believe that the current observation is not a reason for the given target value. However, his expectation that the observed value for E_i makes his expectations more likely is not justified. We give this conclusion to the user in the following way; *The observation for E_i is indeed not an important finding for t. However, it is also not a reason for t', because other values for E_i make t' more likely.*

In the last situation E_i is part of both sufficient and counterfactual sets. Because E_i is in a sufficient set it follows that this observation was important and that the most probable value will be less likely if this observation was not given. That E_i is in a counterfactual set indicates that other values for E_i make the expected value of the target more likely. It follows that the user has the wrong expectations of how E_i influences the target. This can be given to the user in the following way. *You expectations of the observation for E_i were not correct. The observed value for E_i makes t more likely and another value for E_i makes t' more likely.*

S	Based on the given evidence, what is your expected value for *Acccident*?
U	I expect a moderate accident.
S	Based on the given evidence, it is found that *severe* is the most likely value for *Accident*. Do you prefer understandable or convincing explanations for this finding?
U	I prefer understandable explanations.
S	We found the following sufficient and counterfactual explanations: 1. If only $Age = adolescent, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence. If $Model = normal$ was observed instead of the actual values and all other values would stay the same, *moderate* would be the most probable value for Accident 2. If only $Age = adolescent, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence. If $Model = old$ was observed instead of the actual values and all other values would stay the same, *moderate* would be the most probable value for Accident. 3. If only $Mileage = 2000, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence. If $Model = normal$ was observed instead of the actual value and all other values would stay the same, *moderate* would be the most probable value for Accident. Do you want to see more explanations?
U	No.
S	On what observation is your expectation for a moderate accident based?
U	On the observation that I have a cautious driving style.
S	You probable have the right expectations of how E_i influences the target. However, this influence was suppressed by the other observations.

Table 7: An example of how the explanations can be given to a user, where U gives the statements of the user and S gives the statements of the system.

9.4 A conversation with the user

In this section we show an example of how the explanation is given to the user in practice. We use the example from Section 8.5, where a BN was used to determine how severe an accident for a driver would be. Table 7 shows how the explanations and additional information can be given to the user during conversation.

First the user is asked for his expectations for the target value based on the evidence. After he stated this expectation, the most probable value is computed and the user is asked to state his preference for understandable or convincing explanations. In the example the user prefers understandable explanations. Based on this preference we compute the explanations. Because some evidence is excluded from the lattice, we use Proposition 9.2 to simplify the algorithm. Not all explanations will be found now, however the most understandable will be. All explanations that are found have the same conciseness and uniformity. Three explanations are picked randomly and Templates 9.1 and 9.3 are used to present them to the user.

After the explanations are given, the user states that he does not want to see more explanations. To be able to give the user some additional information, we ask why he had the expected value for the target. Based on his answer we are able to derive that the user has the right expectations for this finding, but the influence of this finding is suppressed by the other observations. This conclusion is conveyed to the user.

10 Conclusion

In Section 3 a literature research about different explanation methods for AI systems and more specifically for BNs is described. From that research the following conclusions were reached. First of all, we saw that people generally

prefer a contrastive explanation, where one value of the target variable of a system is contrasted against another value. However, these kind of explanations for AI systems are relatively unexplored. Secondly we saw that in recent years it is explored how counterfactual statements could be used to explain how a minimal change to input variables could change the output of a system. This is however not yet explored for Bayesian networks specifically. This lead us to the following research question; How can a contrastive, counterfactual explanation for a target variable of a Bayesian network be obtained and conveyed to the user?

We answered this question by splitting it into three subquestions. With the first subquestion we provided a definition of a contrastive, counterfactual explanation. To give a useful definition we first explored what elements such an explanation should contain. We decided that the explanation should always have a target. This target has one value that is the most probable given the evidence in a system and one value that was expected by the user. It was concluded that a contrastive, counterfactual explanation exists of the following two sets. First of all, it contains a sufficient explanation. Given the observations for the variables in this explanation, the other variables could have an arbitrary value without changing the most probable value of the target. Secondly, the contrastive, counterfactual explanation should have a counterfactual explanation. This explanation states what minimal changes should be applied to the observations for the target to result in the expected value of the user. The combination of the sufficient and the counterfactual explanation give a contrast between the expected and most probable value of the target. A formal definition of the contrastive, counterfactual explanation is given by Definition 4.4. This definition does not use any properties of a BN. It follows that the given definition is applicable to any kind of system that uses different observations to compute a most probable value for a target.

After we defined the explanation, we explored for the next subquestion how this could be computed from a BN. We started with a naive approach for evidence that was binary valued. This approach gave valid explanations, however we concluded that it was not efficient. As a result we gave Definition 6.1 of an enhanced subset lattice as a framework of finding all explanations. We defined how all sufficient explanations are found by a breadth-first search while dynamically constructing this lattice. The result is given in Algorithm 3. Algorithm 4 defines how all counterfactual explanations are found in a similar way. We gave an example of how both algorithms can be combined to find all explanations in the lattice with one breadth-first search. We concluded that in the worst-case scenario a probability needed to be computed for all 2^n subsets in the lattice, with n the size of the evidence set. We also saw however, that this worst-case scenario was not likely to occur often. The given approach could be extended to also be applicable in case the evidence is not binary valued. For this approach to work k^m probabilities need to be computed for each node in the lattice, with k the maximum number of possible values for a variable and m the size of the inverse of the subset.

To come up with a more feasible approach for non-binary valued evidence, we explored how monotonicity could be used to compute all explanations. In Section 7 we gave two different definitions for monotonicity and derived several propositions about the inclusion of observations in sufficient or counterfactual sets if a monotonicity relation between the evidence and the target was assumed. From these propositions it was concluded that monotonicity in mode is the more useful concept of the two for computing all contrastive, counterfactual explanations. We proved with several propositions that some evidence variables could never be part of a sufficient or counterfactual set. Based on this observation we decided to exclude those evidence variables when building a subset lattice. We also excluded some additional evidence variables to enforce some useful properties in the lattice. Definition 8.5 gives the definition for the monotonicity enhanced subset lattice for evidence that is monotone in mode with the target. Based on this definition Algorithm 5 performs a breadth-first search through the lattice to find all sufficient sets. However, because some evidence variables were excluded from it some additional computations are needed to find all sets. How these computations are done is given in Algorithm 6. Algorithm 7 gives how counterfactual explanations can be found in the lattice with one breadth-first search. Again does this algorithm not find all counterfactual explanations, because the sets with evidence variables that were excluded from the lattice are not found. Algorithm 8 gives how the Algorithm 7 is extended to also find those counterfactual explanations.

In conclusion, we described two feasible ways of how a contrastive, counterfactual explanation can be computed from a BN. The first method only works for binary valued evidence. The second method only works for evidence that is monotone in mode with the target.

The last subquestion focused on how the explanations can be given to the user. First we defined two different ways of how a selection can be made of the explanations. The first method selects the explanations that are easily understandable, where an explanation is considered more understandable if it is more concise and more uniform. The second method selects explanations that are more convincing, where we define an explanation to be more convincing if the probability of the target given the sufficient set, the probability of observing the counterfactual explanation and the probability of the expected value of the target are all high. Different templates of how the sufficient and counterfactual explanations are given to the user in a textual way are presented. We also defined

how a user could make a selection of the evidence variables to give counterfactual explanations with only evidence variables he has an interest in. At last we demonstrated how some additional information could be provided to the user based on the observations he gave as his reason for the expected target value.

11 Discussion

In this section we discuss different aspects of the research that could be improved or that gives rise to additional research. First of all we note that the algorithms we present do not use any characteristics or properties of a BN. This is only done when a probability is computed and in that case a standard algorithm is used. Perhaps these algorithms could be adjusted in a way to more efficiently compute sufficient or counterfactual sets. Or perhaps the structure of the graph of the BN could be used to improve the definition of the enhanced subset lattice. Future research could explore this. However, because no properties of the BN are used, it follows that the presented algorithms can be easily adjusted to be used by other AI systems. The only requirement for those systems is that they use a set of evidence to compute the most probable value for a target.

The algorithms presented in Section 8 use the assumption that the evidence is monotone in mode with the target to find all explanations. L. Van der Gaag et al. (2004) present an approximate algorithm for deciding if evidence is monotone in *distribution*. How an assignment lattice can be used to identify evidence that violates the properties of monotonicity in *distribution* is later explored by L. C. Van der Gaag et al. (2006). However, there is not yet found a way of efficiently verifying if evidence is monotone in mode in a BN. It follows that the algorithms presented in this thesis that build on the monotonicity assumption can not be applied before such a method is found. However, even if there is not an efficient method of verifying monotonicity, it could be explored if Algorithms 5 and 8 are useful as approximate algorithms. For example, if there is a strong indication that the evidence in a BN is monotone in mode with the target, we can then use Algorithms 5 and 8 to find all explanations. The algorithms would then not necessarily find all explanations and they would maybe return explanations that are not minimal. However, the explanations that are found can still be considered useful by the user. Future research should indicate if this is indeed the case.

Both Algorithms 5 and 8 need to compute 2^n modes to find all explanations for an evidence set of size n in the worst case. We expect that this worst case does not occur often. However, we can not yet make any useful statements about the efficiency of the algorithms in practice. Because all explanations are found by a breadth-first search starting at the top of the lattice, it follows that the algorithms terminate earlier if the sufficient and counterfactual sets are found in nodes closer to the top of the lattice. Experimenting with several BNs and different sizes of evidence sets could give us an indication of where in the lattice the explanations are usually located. If it turns out that most explanations are found in nodes closer to the bottom of the lattice, it could be explored if starting the search at the bottom results in a more efficient search.

In case the evidence in a BN is not binary valued and does not have a monotone relation with the target, finding all explanations can not be done in an efficient way. It follows from the definition of the sufficient explanation that a probability needs to be computed for all possible value configurations of the evidence not in the explanation. If this definition is adjusted, it could be possible to compute sufficient explanations more efficiently. For example, given a set of evidence $\mathbf{E}$ and most probable value t for T we can say that a set $\mathbf{S}$ is sufficient if t is the most probable value for the target given *only* these observations for $\mathbf{S}$. With this definition it is not enforced to compute multiple probabilities before it can be decided that a set is sufficient. However, this definition also gives less information to the user.

The definition for a contrastive, counterfactual explanation as given in this thesis is justified by a literature research, however it is not yet tested if this explanation is indeed helpful for the users of a system. Experimenting with different subjects could tell us the following. First of all it can show if contrastive, counterfactual explanations do indeed help the user to better understand the domain of the BN. Secondly we can determine if the user has more trust in the system after a contrastive, counterfactual explanation is given. At last we should determine if the given templates are easily understandable for the user or that the way in which the explanation is presented to the user should be improved.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Andersen, S. K., Olesen, K. G., Jensen, F. V., & Jensen, F. (1989). HUGIN-a Shell for Building Bayesian Belief Universes for Expert Systems. In *Eleventh International Joint Conference on Artificial Intelligence* (Vol. 89, pp. 1080–1085).

Baker, M., & Boult, T. E. (2013). Pruning Bayesian networks for efficient computation. *arXiv preprint arXiv:1304.1112*.

Binder, J., Koller, D., Russell, S., & Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning*, *29*(2-3), 213–244.

Druzdzel, M. J. (1993). *Probabilistic reasoning in decision support systems: from computation to common sense* (PhD thesis). Carnegie Institute of Technology.

Etchells, T. A., & Lisboa, P. J. (2006). Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach. *IEEE transactions on neural networks*, *17*(2), 374–384.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In *Proceedings of the 36th international conference on machine learning* (p. 2376–2384).

Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018). Interpretable credit application predictions with counterfactual explanations. In *Workshop on challenges and opportunities for ai in financial services: the impact of fairness, explainability, accuracy, and privacy,*.

Grätzer, G. (2011). *Lattice theory: foundation*. Springer Science & Business Media.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497–530.

Haddawy, P., Jacobson, J., & Kahn Jr, C. E. (1997). BANTER: a Bayesian network tutoring shell. *Artificial Intelligence in Medicine*, *10*(2), 177–200.

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian networks and decision graphs* (2nd ed.). Springer Science & Business Media.

Kyrimi, E., et al. (2019). *Bayesian Networks for Clinical Decision Making: Support, Assurance, Trust* (PhD thesis). Queen Mary University of London.

Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, *17*(2), 107–127.

Lacave, C., Oniśko, A., & Díez, F. J. (2006). Use of Elvira's explanation facility for debugging probabilistic expert systems. *Knowledge-Based Systems*, *19*(8), 730–738.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *50*(2), 157–194.

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing* (pp. 195–204).

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319–330.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Oxford English dictionary [Explanation]. (n.d.). Oxford University Press. Retrieved 21 February 2020, from `https://www.lexico.com/definition/explanation`

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach* (3rd ed.). Pearson Education Limited.

Santos Jr, E. (1991). On the generation of alternative explanations with implications for belief revision. In *Uncertainty proceedings 1991* (pp. 339–347). Elsevier.

Setiono, R., & Leow, W. K. (2000). FERNN: An algorithm for fast extraction of rules from neural networks. *Applied Intelligence*, *12*(1-2), 15–25.

Shih, A., Choi, A., & Darwiche, A. (2018). A symbolic approach to explaining Bayesian network classifiers. *In IJCAI*, 5103–5111.

Shimony, S. E. (1991). *A Probabilistic Framework for Explanation* (PhD thesis). Brown University.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647–665.

Suermondt, H. J. (1992). *Explanation in Bayesian belief networks.* (PhD thesis). Stanford University.

Suermondt, H. J., & Cooper, G. F. (1993). An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, *26*(3), 242–254.

Teach, R. L., & Shortliffe, E. H. (1981). An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, *14*(6), 542–558.

Timmer, S. T. (2017). *Designing and understanding forensic Bayesian networks using argumentation* (PhD thesis). Utrecht University.

Van der Gaag, L., Bodlaender, H. L., & Feelders, A. (2004). Monotonicity in bayesian networks. In *20th Conference on Uncertainty in Artificial Intelligence* (p. 569–576).

Van der Gaag, L. C., Renooij, S., & Geenen, P. L. (2006). Lattices for studying monotonicity of bayesian networks. In *Probabilistic graphical models* (pp. 99–106).

Van Leersum, J. (2015). *Explaining the reasoning of Bayesian networks with intermediate nodes and clusters* (Master's thesis). Utrecht University.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Yap, G.-E., Tan, A.-H., & Pang, H.-H. (2008). Explaining inferences in Bayesian networks. *Applied Intelligence*, *29*(3), 263–278.

Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). DeepRED– Rule extraction from deep neural networks. In *International conference on discovery science* (pp. 457–473).