Lecture 11
Independence and Bayesian Networks

Alice Gao

June 7, 2021

Contents

1 Learning Goals **2**

2 Unconditional and Conditional Independence **2**
2.1 Definitions 2
2.2 Deriving a compact representation 3

3 Examples of Bayesian Networks **7**

4 Why Bayesian Networks? **12**

5 Components of a Bayesian network **13**

6 Representing the joint distribution **15**

7 Independence relationships in three key structures **17**

1 Learning Goals

By the end of the lecture, you should be able to

- Given a probabilistic model, determine if two variables are unconditionally independent.
- Given a probabilistic model, determine if two variables are conditionally independent given a third variable.
- Give examples of deriving a compact representation of a joint distribution by using independence and/or conditional independence assumptions.
- Describe components of a Bayesian network.
- Compute a joint probability given a Bayesian network.
- Given a Bayesian network, determine if two variables are independent or conditionally independent given a third variable.

2 Unconditional and Conditional Independence

This will be a short review of two important concepts in probability theory: unconditional independence and conditional independence. First we'll go over the formal definitions of these two probability assumptions, then we'll go over why it is important to understand these definitions for constructing a Bayesian network.

2.1 Definitions

Definition ((Unconditional) independence)**.** X and Y are (unconditionally) independent if and only if

$$P(X|Y) = P(X)$$
$$P(Y|X) = P(Y)$$
$$P(X \wedge Y) = P(X)P(Y)$$

Often, when referring to this definition, we will simply say X and Y are independent.

Intuitively, this definition means that learning Y does *not* influence your belief about X (and symmetrically for learning X). This is what the first two formulas are saying.

The third formula is saying the joint probability of X and Y is the product of their marginal or prior probabilities. This is equivalent to the first two formulas together. To see this equivalence, you can multiply the first formula by $P(Y)$ or the second formula by $P(X)$. For example, multiplying the second formula by $P(X)$ gives

$$P(X \wedge Y) = P(X)P(Y|X) = P(X)P(Y)$$

which is exactly the third formula.

Note that the third formula says that we only need two values to specify the joint distribution of X and Y (their prior probabilities), rather than the usual four values (one for each combination of truth values for X and Y). We'll see how this can help us find a more compact representation of a probability distribution.

Definition (Conditional independence)**.** X and Y are conditionally independent given Z if and only if

$$P(X|Y \wedge Z) = P(X|Z)$$
$$P(Y|X \wedge Z) = P(Y|Z)$$
$$P(X \wedge Y|Z) = P(X|Z)P(Y|Z)$$

Intuitively, this definition means that learning Y does *not* influence your belief about X if you already know Z (and symmetrically for learning X).

The third formula is very similar to the one in the previous definition, but the probabilities are conditional on Z. Like with the previous definition, we can derive the third formula from the first two. This is left as an exercise for you—the derivation is a little more interesting than before because it involves some conditional probabilities.

Having seen these two definitions, you may be wondering: "These definitions sound similar. Do they have some inherent relationship? If two variables are independent, are they conditionally independent, and vice versa?" Unfortunately, these definitions have no inherent relationship. If we know one is satisfied, it tells us nothing about whether the other is satisfied.

2.2 Deriving a compact representation

Now we'll use these definitions to see how we can derive a compact representation of a probability distribution and how we can verify the independence relationships between some variables given a probability distribution.

Problem: Consider a model with three Boolean random variables, A, B, C.

1. What is the minimum number of probabilities required to specify the joint distribution?

2. Assume that A, B, and C are independent. What is the minimum number of probabilities required to specify the joint distribution?

Solution:

1. In general, given n Boolean random variables, we need $2^n - 1$ probabilities to specify the joint distribution. In this case, we would need 7 probabilities.

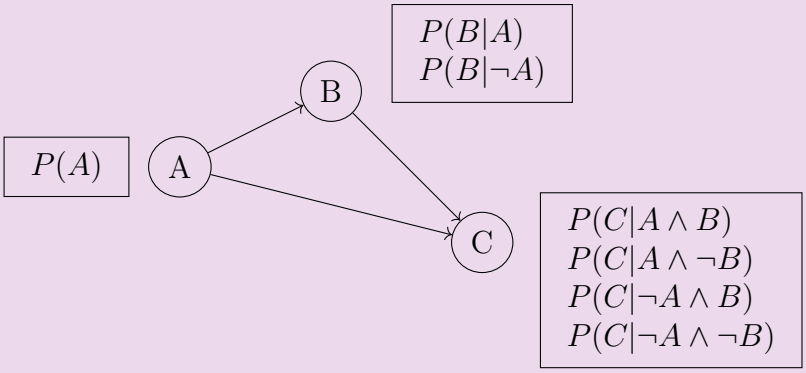
   This is because we have to specify probabilities like, in this case, $P(A \wedge B \wedge C)$.

Each variable has two possibilities (true or false), but because all the probabilities in the distribution sum to 1, we can recover the last probability by subtracting the others from 1.

You can also think about this problem in another way, which may sound strange now but may help connect this topic to Bayesian networks later. If you remember the chain rule, we can expand this joint probability as

$$P(A \wedge B \wedge C) = P(A) * P(B|A) * P(C|A \wedge B).$$

We can represent this graphically using nodes for the variables and a directed edge from one variable to another if the latter is conditional on the former.



In the diagram, the probabilities we'd need to specify for that node in order to determine the joint distribution is written beside each node. You can see for A, we only need $P(A)$ as we can find $P(\neg A)$ by computing $1 - P(A)$. Similarly, for B we only need to specify the probabilities where B is true; there are two such probabilities. Finally, for C we only need to specify the probabilities where C is true; there are four such probabilities.

Based on this picture, we need to specify at least $1 + 2 + 4 = 7$ probabilities to recover the joint distribution.

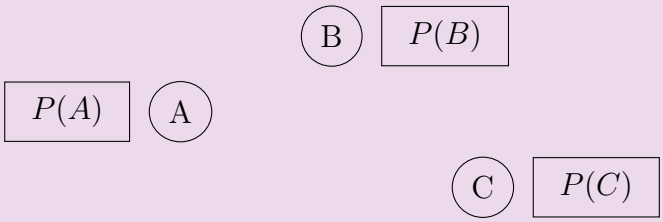
2. In this case, since the variables are independent, the joint probability becomes

$$P(A \wedge B \wedge C) = P(A) * P(B) * P(C).$$

You can see independence at work by comparing the factors: we have $P(B|A)$ become $P(B)$ and $P(C|A \wedge B)$ become $P(C)$.

The result is we only need to specify three probabilities here (one for each variable).

You can again reason about this using a picture. Updating our graph, we have no more edges. As a result, we only need to specify a single probability for each node in order to specify the joint distribution.



Problem: Consider a model with three Boolean random variables, A, B, C.

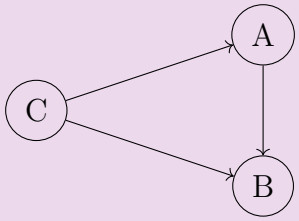
1. What is the minimum number of probabilities required to specify the joint distribution?

2. Assume that A and B are conditionally independent given C. What is the minimum number of probabilities required to specify the joint distribution?

Solution:

1. See the previous problem.

2. Here is another graphical model to represent the joint distribution. Notice that the directions of the arrows is a bit different from the previous problem. This is because this model is more convenient for explaining the situation here.



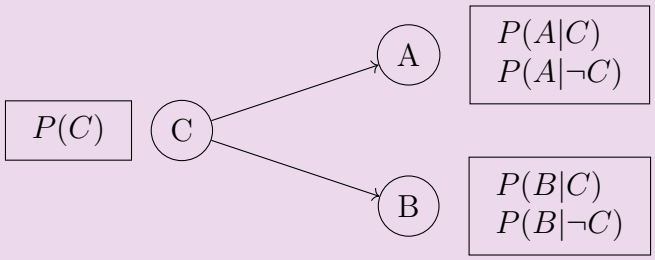
Mathematically, this model corresponds to this expansion:

$$P(A \wedge B \wedge C) = P(C) * P(A|C) * P(B|C \wedge A)$$

Now let's apply the conditional independence assumption. Under this assumption, $P(B|C \wedge A) = P(B|C)$: since A and B are conditionally independent, knowing A does not influence our belief about B if we already know C.

$$= P(C) * P(A|C) * P(B|C)$$

Graphically, this means we can delete the edge from A to B since the value of B does not depend on A. After deleting this edge, we can again count the number of probabilities we need to recover the joint distribution. For C, we only need 1. For each of A and B, we only need 2.



We need at least $1 + 2 + 2 = 5$ probabilities here.

Though not formally defined yet, the graphical models in diagrams above are essentially Bayesian networks. The reason for using both a mathematical derivation and a graphical representation is to show you different ways of thinking about the same problem. The important message is that if we know some independence assumptions or conditional independence assumptions about these variables, that allows us to use fewer probabilities to represent the same joint distribution.

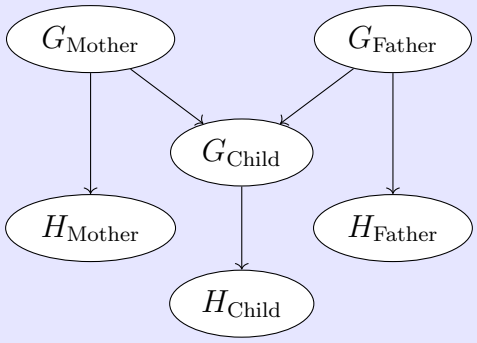
3 Examples of Bayesian Networks

After doing quite a bit of review on probability theory, we're finally ready for a new unit on probabilistic inference and reasoning using Bayesian networks.

Let's first consider a few examples of the kinds of problems we might want to model using Bayesian networks. You may not have seen many examples of Bayesian networks in your daily life, but hopefully once you see these examples, you will realize they are natural models for certain scenarios.

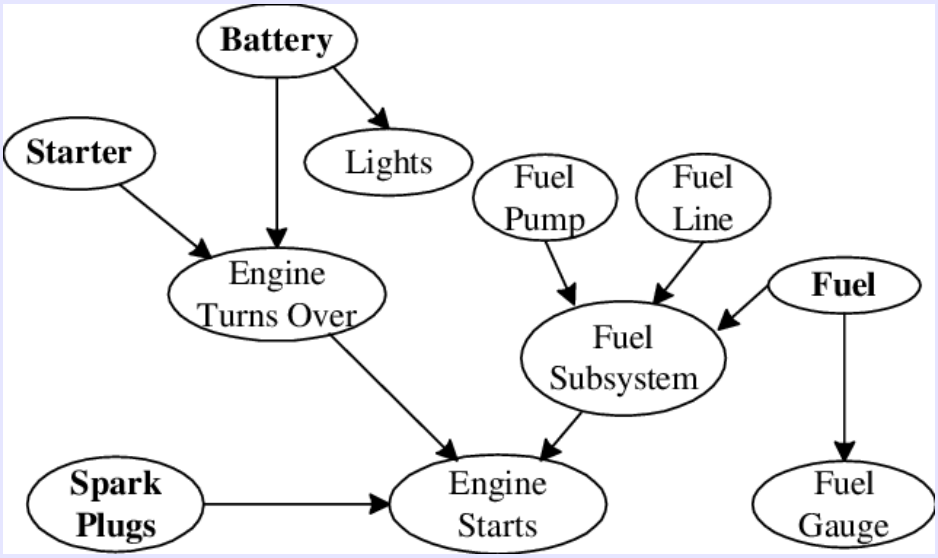
Example: This first example is about genetics and modelling the inheritance of handedness: whether someone is left- or right-handed.

We have nodes G_{Mother} and G_{Father} modelling the handedness genes of the parents. Given the handedness in the parents, those genes will decide on the genes in the child, represented by G_{Child}. However, whether and how these genes will be expressed in the individual might differ from what's expected. So, we have additional nodes H_{Mother}, H_{father}, and H_{Child} modelling how the genes are expressed in the mother, father, and child. Finally, we have (directed) edges representing the nodes' relationships.



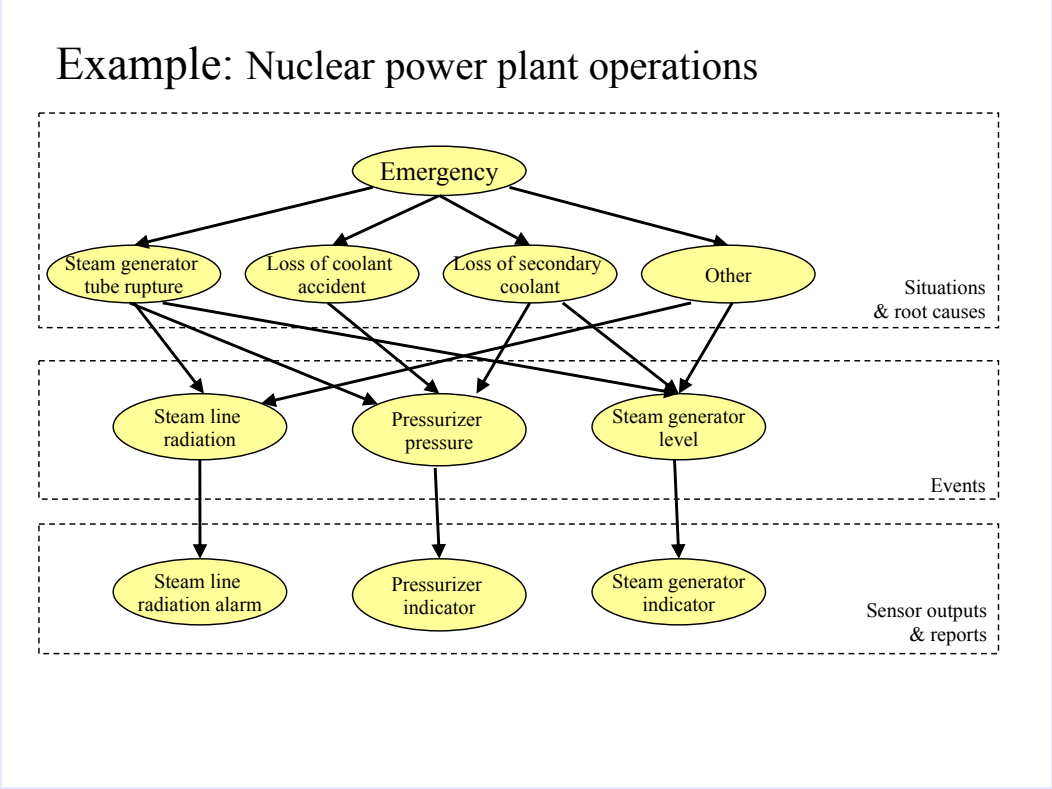
Example: This is an example of a car diagnostic network. You will see that a lot of the Bayesian network examples are about diagnosing problems in certain scenarios.

In this scenario, we may have problems in a particular part of the car, and that may be caused by many different reasons. The battery might cause the lights to have different problems or it may cause the engine to have problems, and that in turn may affect whether or not the engine can start.



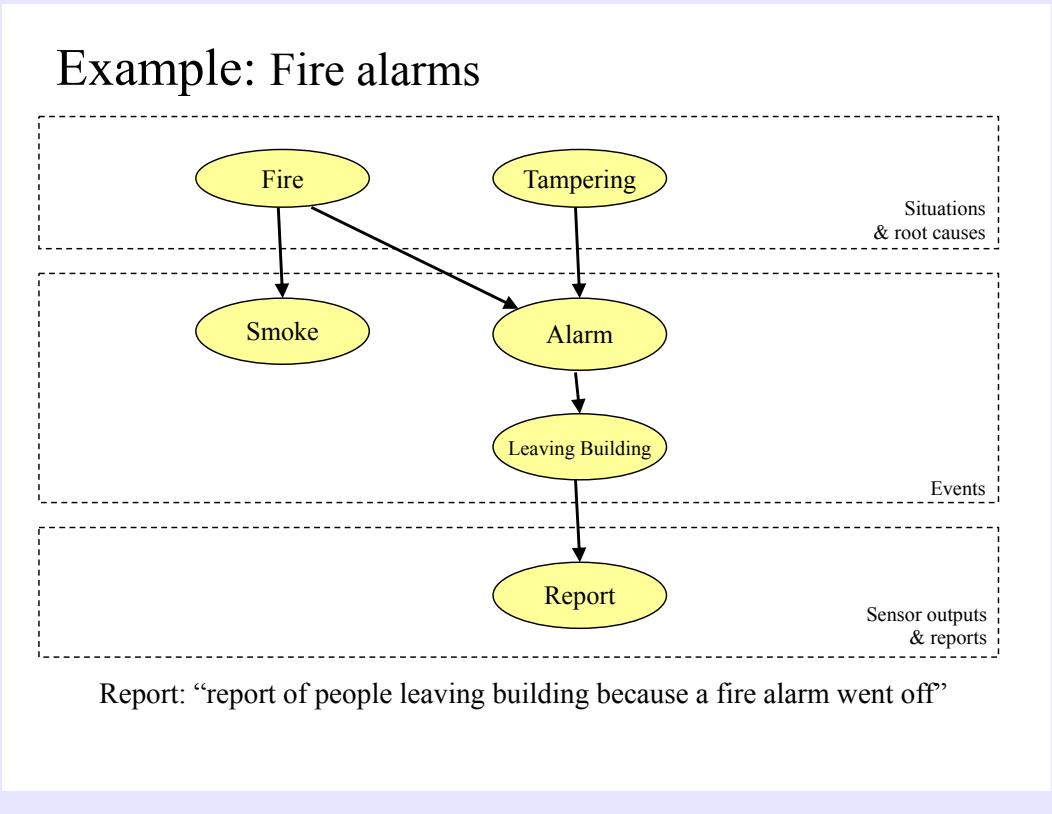
Example: This is a network on diagnosing the problems in a nuclear power plant. This network has some interesting structure. We have three tiers in this network: the top tier represents situations and root causes, the middle tier represents the events we might have because of these root causes, and the bottom tier represents the sensor outputs and reports that may alert us of the events.

A lot of real-world scenarios are like this. There are a lot of root causes, the items in the top tier that we never really get to observe for ourselves. Then there are events happening in the middle, which we again often do not observe directly. Instead, we only observe the output from sensors and alarms in the bottom tier, which allows us to know that something is happening.



Example: This is a much simpler scenario where we have a fire alarm going off in the building. This network is saying that an actual fire could cause smoke or an alarm, and then it could cause people to leave the building, and finally that would cause others to report to the authorities that something might be happening. However, another possible cause for the alarm could be somebody tampering with the alarm.

Our Holmes scenario is quite similar to this. There, we have an alarm and two possible causes for the alarm—burglary or earthquake—but we do not directly observe the alarm. Instead, we only observe calls from Dr. Watson or Mrs. Gibbon notifying Mr. Holmes that the alarm might be going off and that maybe a burglary is happening.

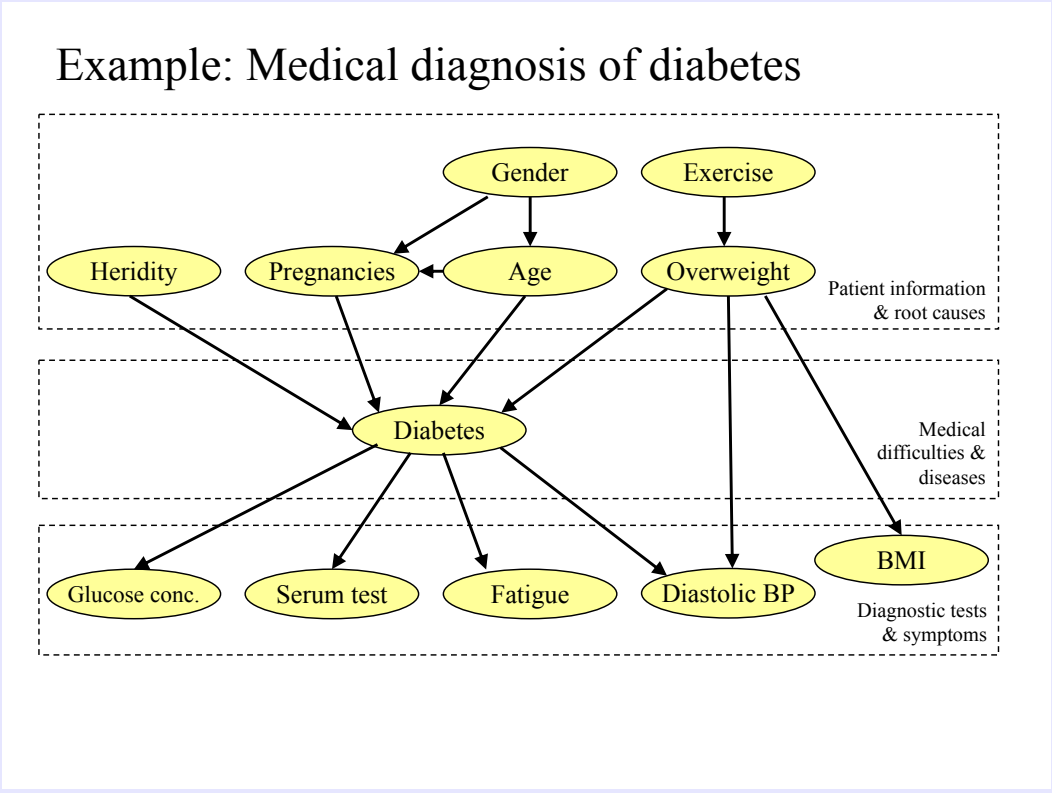


Report: "report of people leaving building because a fire alarm went off"

Example: Another prominent application of Bayesian networks is in a medical diagnosis scenario. Recalled that we often know certain diseases will have certain symptoms, but we never directly observe whether somebody has a disease. We only really observe the symptoms and we want to infer whether they have a particular disease.

So, a lot of medical diagnosis scenarios are like that: we have some underlying disease (shown here in the middle), this disease could itself have underlying causes (shown here at the top), and finally the disease could cause certain symptoms (shown here at the bottom). As doctors or medical professionals, we only really observe the results of diagnostic tests and symptoms, and we use this information to try to infer whether somebody has a disease.

In this scenario, it's interesting that we may also observe things from the top layer. In addition to symptoms and tests, we could also observe patient information to inform our diagnosis: the patient's age, weight, medical history, genetics, and so on.



4 Why Bayesian Networks?

During the review of probability theory, we introduced the Holmes scenario. Mr. Holmes is worried about burglary happening at his home so he installs an alarm, but he relies on his neighbours, Dr. Watson and Mrs. Gibbon, to call him to tell him that the alarm might be going off. He also learned that the alarm might be triggered by an earthquake, but if an earthquake is happening, it's very likely that there will be a radio report of the earthquake.

We talked about how to model this scenario using random variables and probabilities. We defined six random variables, for Earthquake, Radio, Burglary, Alarm, Watson, and Gibbon. There are a total of $2^6 = 64$ probabilities in the joint distribution, so we need to specify a minimum of 63 probabilities in order to represent the distribution.

Also from the review, you should remember that knowing the joint distribution is very powerful: knowing it, we have enough information to do any probabilistic inference (finding prior or conditional probabilities) in the network.

So why don't we just write down this giant joint distribution and use it to do inference and call it a day? Why would we use Bayesian networks instead?

There are two main reasons.

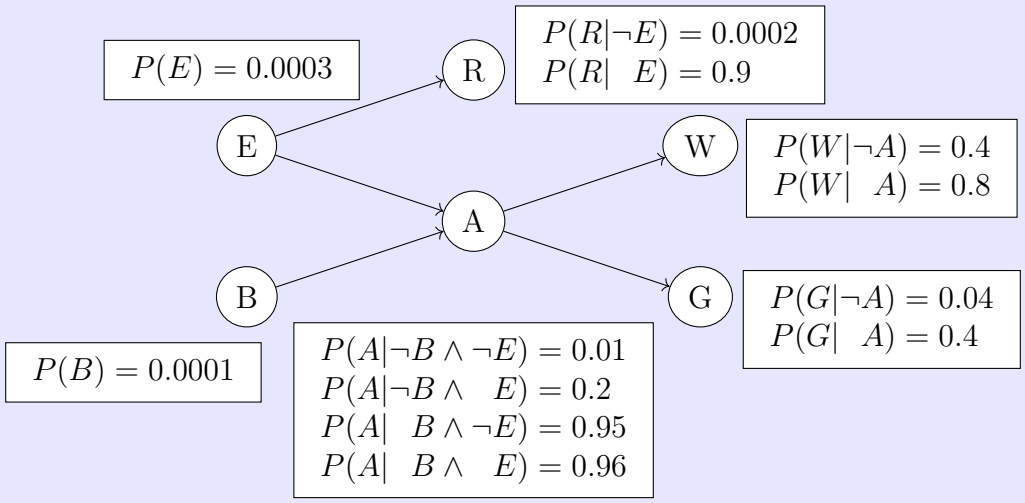
- It quickly becomes intractable to compute probabilities using the joint distribution, since the number of probabilities (2^n) is exponential in the number of variables (n).
- It is unnatural and tedious to specify all the probabilities — thinking about the joint probability of even just our six variables for the Holmes scenario is unintuitive.

To address these concerns, we'd like a more compact and efficient way of representing the joint distribution. These are the most important reasons leading us to Bayesian networks.

In short, a Bayesian network

- is a compact representation of the joint distribution, and
- takes advantage of the unconditional and conditional independence among the variables.

Example: Here is a possible Bayesian network for the Holmes scenario.



Let's first look at our savings in terms of the number of probabilities. Remember that to specify the full joint distribution, at least $2^6 - 1 = 63$ probabilities are required.

Given this Bayesian network, only $1+1+2+2+2+4 = 12$ probabilities are required to specify the joint distribution.

This is a much more compact representation of the distribution, and we'll see that this could actually represent the same thing as the joint distribution.

Another important point to realize is that there are many possible Bayesian networks that can represent the same scenario. This example should be the best network in terms of minimizing the total number of probabilities that we need, but there are other possible networks. Depending on how we order the nodes and draw the directed edges, we might come up with a slightly different network.

Later, we will discuss how we can construct a Bayesian network given a probability distribution. Then, we'll talk about how this construction process affects the structure of the network — in other words, how we can come up with a network with the fewest links so that the number of probabilities required to define the network can be minimized.

5 Components of a Bayesian network

A Bayesian network is a directed acyclic graph.

- Each node corresponds to a random variable.
- X is a parent of Y if there is an arrow from node X to node Y. X is an ancestor of Z and Z is a descendent of X if there is some path from X to Z.
- Each node X_i has a conditional probability distribution $P(X_i|\text{Parents}(X_i))$ that quan-

tifies the effect of the parents on the node. A node with no parents will only require a prior (unconditional) probability.

6 Representing the joint distribution

What does a Bayesian network mean? And how should we interpret it? We will discuss two ways to understand Bayesian networks: first, as a representation of the joint probability distribution; second, as an encoding of the conditional independence assumptions.

Operationally, recovering the joint distribution from a Bayesian network is very simple. We can compute each joint probability using the following formula:

$$P(X_n \wedge \cdots \wedge X_1) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)).$$

In words, the joint probability is the product of the conditional probability of each variable given its parents.

Let's look at some examples. We will go through one example together with calculations, then a question for you to work through the calculations.

Problem: What is the probability that

- the alarm has sounded,
- neither a burglary nor an earthquake has occurred,
- both Watson and Gibbon call and say they hear the alarm, and
- there is no radio report of an earthquake?

Solution: This is a scenario where both Watson and Gibbon send false alarms. Would you expect the probability to be large or to be small?

Let's compute the probability of this scenario happening. The variables are ordered based on the Bayesian network so it is easier to follow.

$$\begin{aligned} P(\text{scenario}) &= P(\neg B \wedge \neg E \wedge A \wedge \neg R \wedge W \wedge G) \\ &= P(\neg B) \cdot P(\neg E) \cdot P(A | \neg B \wedge \neg E) \cdot P(\neg R | \neg E) \cdot P(W|A) \cdot P(G|A) \\ &= (1 - 0.0001)(1 - 0.0003)(0.01)(1 - 0.0002)(0.4)(0.8) \\ &= 0.0032. \end{aligned}$$

It turns out the probability is quite small. This is a good thing! This means that the sensors in this model (the alarm, Watson, and Gibbon) are fairly reliable, with only a small probability that the sensors all go off when nothing is actually happening.

Try the next practice problem on your own.

Problem: What is the probability that

- **neither** a burglary **nor** an earthquake has occurred,
- the alarm has **not** sounded,
- **neither** Watson **nor** Gibbon is calling, and
- there is **no** radio report of an earthquake?

(A) 0.5699

(B) 0.6699

(C) 0.7699

(D) 0.8699

(E) 0.9699

Solution: Here, we have

$$\begin{aligned} P(\text{scenario}) &= (1-0.0001)(1-0.0003)(1-0.01)(1-0.4)(1-0.04)(1-0.0002) \\ &= 0.5699. \end{aligned}$$

The correct answer is (A).

In this scenario, nothing "exciting" is happening in the world. ("No news is good news!") We hope for this probability to be somewhat high.

Before moving on, let's come back to the formula:

$$P(X_n \wedge \cdots \wedge X_1) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)).$$

We know how to use this formula, but we haven't talk about why this formula makes sense. In the next subsection, we'll see why it makes sense to calculate the joint probabilities in this way.

For now, think about this question. We previously talked about how to calculate a joint probability using the chain rule. Try to apply the chain rule to the joint probability on the left-hand side of the formula and compare what you get with the expression on the right-hand side—term-by-term, especially. Pick a particular variable and compare the term you get from the chain rule with the term you get from the formula. What do you notice? What's the difference between the two terms, and what does that tell you about the Bayesian network?

7 Independence relationships in three key structures

To see how a Bayesian network can encode independence relationships among the random variables, let's first look at three examples from the Holmes scenario. Each example is fairly small, only containing three random variables, but they are chosen very carefully because they represent the most basic and fundamental relationships we can see in a Bayesian network. There is an intuitive explanation for each example and, additionally, more formal explanations for two of the examples.

In this first key structure, we consider a chain of three nodes.

Problem: Are Burglary and Watson independent?

(A) Yes (B) No

Solution: The correct answer is no.

We must ask: if we learned Burglary, would our belief of Watson change? Intuitively, yes: if Burglary is happening, then Alarm is more likely to be going off, and Watson is more likely to be calling Holmes.

Problem: Are Burglary and Watson conditionally independent given Alarm?

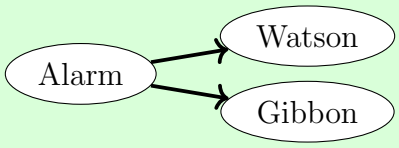
(A) Yes (B) No

Solution: The correct answer is yes.

Watson does not observe Burglary directly; Burglary can only influence Watson through Alarm. If Alarm is known, we have "cut the chain" in the middle. Burglary and Watson cannot affect each other anymore.

In this second key structure, we consider one node with two children.

Problem: Are Watson and Gibbon independent?

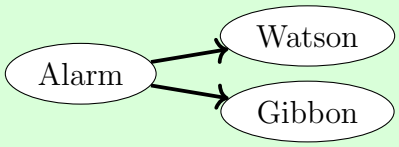


(A) Yes (B) No

Solution: The correct answer is no.

If we learn the value of Watson, does this influence our belief about Gibbon? Yes: if Watson is calling, it is more likely that Alarm is going off, which means that it is more likely that Gibbon is calling.

Problem: Are Watson and Gibbon conditionally independent given Alarm?



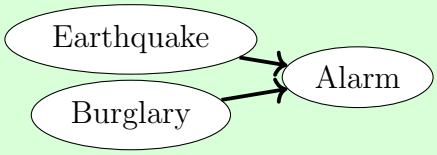
(A) Yes (B) No

Solution: The correct answer is yes.

We can think about Alarm as an event and Watson and Gibbon as noisy sensors for the event. The value of each sensor depends only on the event. If we know whether the event is happening or not, then the two sensors can no longer affect each other.

In this third key structure, we consider one node with two parents.

Problem: Are Earthquake and Burglary independent?



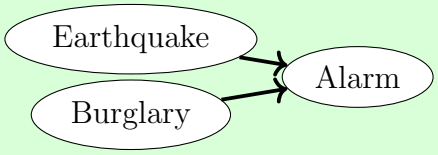
(A) Yes (B) No

Solution: The correct answer is yes.

If we learn whether Earthquake is happening or not, this does not change our belief

about Burglary, and vice versa. (Let's assume that looting is not more frequent during an earthquake!)

Problem: Are Earthquake and Burglary conditionally independent given Alarm?



(A) Yes (B) No

Solution: The correct answer is no.

Suppose that the Alarm is going off. If Earthquake is happening, then it is less likely that the Alarm is caused by Burglary. If Earthquake is not happening, then it is more likely that Burglary is happening and causing the Alarm. This may be called the "explaining away" effect.