Machine Learning Course - CS-433

Graphical Models – Bayes Nets

Dec 20, 2018

Last updated: December 18, 2018

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Outline

Assume that you are given a large set of random variables, call them $X_1, \cdots, X_D$. You might be interested in their relationship. E.g., you might be interested if X_1 is independent of X_2 given lets say X_3. Or perhaps you have a description of the "local" relationships between these random variables and have observed some of them and now you want to know what this tells you about some of the other random variables (inference).

Graphical models is a good framework to answer such questions. As the name suggests, graphical models are models that use a *graphical* depiction of the relationships between random variables. There are quite a few related but distinct such descriptions. To name the most prominent ones, there are *Bayes Nets*, *Markov Random Fields*, and *Factor Graphs*.

In the next few lectures we will look at two of these models, namely Bayes Nets and Factor Graphs. We will discuss their definition and their relationship. We will then see how these graphical representations can be used to answer some basic questions of interest. Finally, we discuss the *sum-product* algorithm, a low-complexity algorithm to compute marginals. It is exact for factor graphs that are trees and it is often a useful approximate for general models. In this first lecture we will discuss Bayes Nets.

Chapter 8 of the book by Christopher Bishop contains most of the material we will be discussing.

Bayes Nets: From Distribution to Graphs

Assume that we have a set of random variables $X_1, \cdots, X_D$ with joint distribution $p(X_1, \cdots, X_D)$. For much of what we will discuss it will not matter if these are discrete or continuous random variables. We will always use the notation $p(\cdot)$ and speak of the probability as if these random variables were discrete. In the continous case just think of $p(\cdot)$ as the density.

A basic representation of a joint distribution that is universally applicable is to use the chain rule to write

$$p(X_1, \cdots, X_D) = p(X_1)p(X_2|X_1) \cdots p(X_D|X_1, \cdots, X_{D-1}). \tag{1}$$

In the above expansion we have used the order $X_1, X_2, \cdots, X_D$ but we could have used any of the $D!$ orders. This degree of freedom will be important. Note that in the chain rule, if we write down the factors in the order we have chosen for the variables, then each factor only contains variables in the conditioning that appear *earlier* in this ordering. This is the defining property of a valid chain rule expansion.

To be concrete, assume that $D = 4$ so that we get

$$p(X_1, X_2, X_3, X_4) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2)p(X_4|X_1, X_2, X_3). \tag{2}$$

There is a very natural graphical representation of this factorization which is shown in Figure 1. Associate one *node* to each of the D random variables (and hence by association to each of the D factors). Label these nodes by the random variable they represent. Draw a *directed* edge from node X_j to node X_i if X_j appears in the conditioning of the term

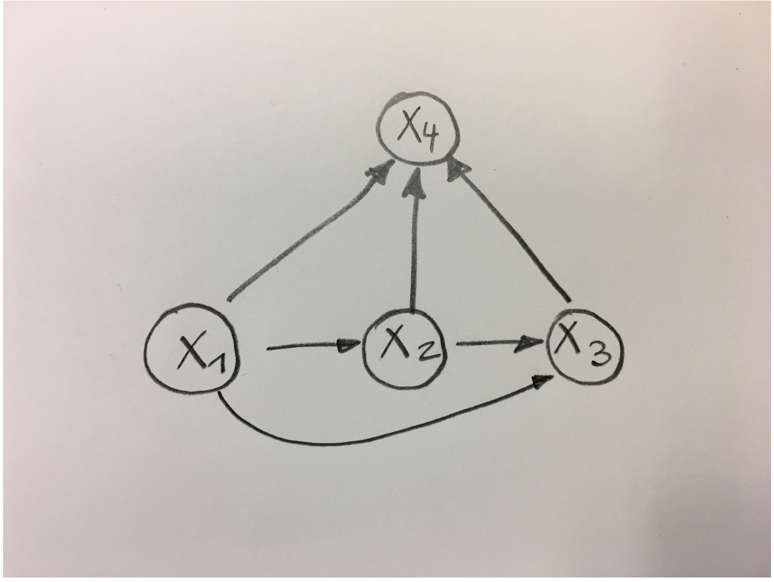

Figure 1: A Bayes net corresponding to the factorization written in (2).

$p(X_i|\cdots)$. We say that X_j is a *parent* of X_i and that X_i is a *child* of X_j.

We can generalize the above procedure by allowing *groups* of random variables at each step rather than a single random variable at a time. But since we can think of a group of random variables as a single random variable (in a larger domain) this is not really more general.

Note that this representation is "universal" and applies to any distribution since so far all we used is the chain rule. So regardless of how we expand the joint distribution we always will get the "same" graph (in the sense that the graph has the same "topology." The representation will become more interesting if some of the edges in this generic graph are missing. E.g., assume that our distribution is such that $p(X_3|X_1, X_2)$ is equal to $p(X_3|X_2)$. In this case we get the graph shown in Figure 2.

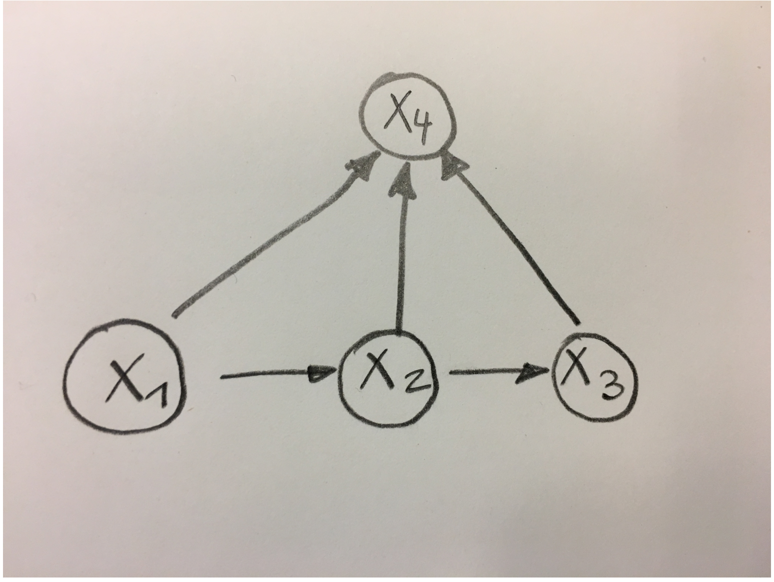

Figure 2: A Bayes net corresponding to the factorization written in (2) with the additional property that $p(X_3|X_1, X_2) = p(X_3|X_2)$.

From Graphs to Distribtions

So far we have gone from a joint distribution to a graph. But assume now that conversely we have a directed graph. When does it correspond to one of these factorizsations? And how can we draw basic conclusions about the relationship of the various random variables?

It is easy to see that not every directed graph corresponds to a valid factorization. E.g., consider the directed graph where X_1 connects to X_2, X_2 connects to X_3 and X_3 connectes to X_1. I.e., we have a cylce on three nodes. This would correspond to the "factorization" $p(X_1 \mid X_2)p(X_2 \mid X_3)p(X_3 \mid X_1)$ which is clearly not valid. I.e., this is not an expression that is the result of expanding a joint distribution in terms of the chain rule and eliminating some conditions.

So a graph corresponds to a valid factorization if there exists an ordering of the variables so that when we write down

the factorization in this ordering, the conditions involved for node/variable lets say X_i only involves variables that appeared earlier in this ordering.
We claim that this is equivalent to saying that the directed graph has to be *acyclic*. This is easy to see. If we have a cycle then just considering those terms on the cycle, it is clear that we cannot write them down in the required form. So being acyclic is necessary. But it is also sufficient.
Assume we are given an acyclic graph. We want to find an appropriate ordering. Note that every acyclic graph has at least one *source*, i.e., a node that has no incoming edges. Why is this true. Start with any node. If this node is a source we are done. Otherwise it has at least one incoming edge. Walk "backwards" along this edge. Now either this new node we reach is a source, in which case we are done, or we repeat. If the graph has D nodes then this procedure cannot involve more than $D-1$ steps since otherwise we have found a cycle.

Conditional independence

Recall that we say that two random variables X and Y are independent if $p(X, Y) = p(X)p(Y)$. And we say that X is independent of Y given Z if $p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$.

Three Basic Examples

Let us now look at three simple graphs, each only involving three variables. This will help us to clarify the important

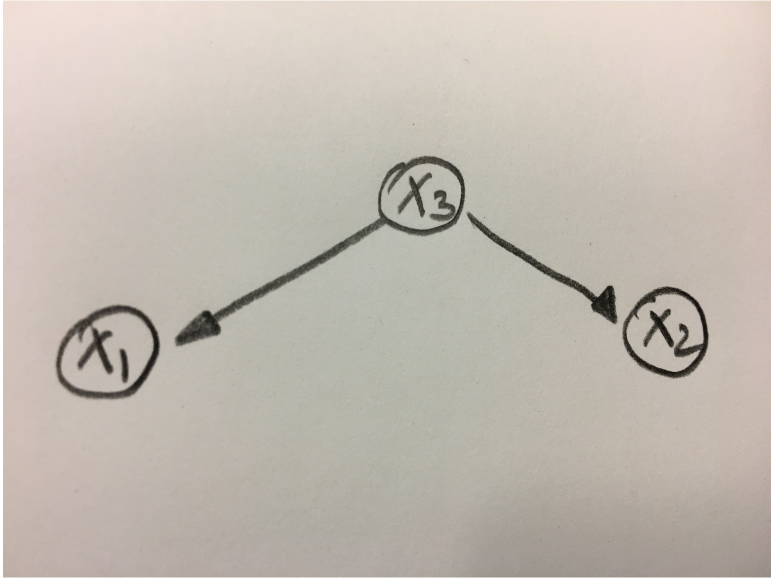

Figure 3: First example: X_3 is *tail-to-tail* with respect to the path from X_1 to X_2. Conditioning leads to independence.

concept of *D-separation.*
We start with the example shown in Figure 3. Note that whenever we talk of a *path* in the sequel we mean an *undirected* path.
This corresponds to the factorization

$$p(X_1, X_2, X_3) = p(X_3)p(X_1|X_3)p(X_2|X_3).$$

If we marginalize out X_3 then, generically, X_1 and X_2 are not independent, i.e., $p(X_1, X_2) \neq p(X_1)p(X_2)$ for this case. But let us look at the conditioned quantity $p(X_1, X_2|X_3)$. We have

$$\begin{aligned} p(X_1, X_2|X_3) &= \frac{p(X_1, X_2, X_3)}{p(X_3)} \\ &= \frac{p(X_3)p(X_1|X_3)p(X_2|X_3)}{p(X_3)} \\ &= p(X_1|X_3)p(X_2|X_3). \end{aligned}$$

So we see that a distribution that has the indicated factorization has the property that X_1 and X_2 are independent *given* X_3. This is sometimes written as

$$X_1 \perp X_2 \mid X_3.$$

Our aim will be to find a simple "graphical" way of determining such (conditional) independence relationships. So let us look at the corresponding Bayes net again.

If you look at the graph you see that X_3 influences both X_1 and X_2. This is why generically these two are not independent. Since X_3 changes in general the distribution of X_1 and X_2, knowing lets say the value of X_1 tells us something about X_3. But this in turn tells us something about the value of X_2.

But, as we just saw, once we condition on the value of X_3 this dependence is gone. Consider the (only) path from X_1 to X_2. It goes via X_3. Note that in this path the two arrows both *point away* from X_3. It is standard terminology to say that X_3 is *tail-to-tail* with respect to this path. As we have just seen, such a tail-to-tail constellation "blocks" the influence going from X_1 to X_2 or vice versa, *assuming that we condition* on X_3.

Let us now move to our next example. It is shown in Figure 4. This corresponds to the factorization

$$p(X_1, X_2, X_3) = p(X_1)p(X_3|X_1)p(X_2|X_3).$$

Now X_1 influences X_3 which in turn influences X_2. Hence X_1 and X_2 are generically not independent.

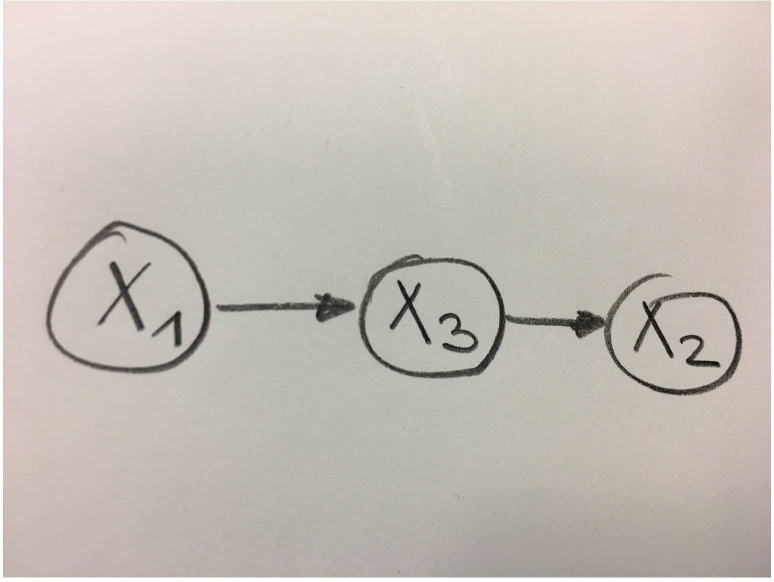

Figure 4: Second example: X_3 is *head-to-tail* with respect to the path from X_1 to X_2. Conditioning leads to independence.

But let us look again at the quantity $p(X_1, X_2|X_3)$. We have

$$\begin{aligned} p(X_1, X_2|X_3) &= \frac{p(X_1, X_2, X_3)}{p(X_3)} \\ &= \frac{p(X_1)p(X_3|X_1)p(X_2|X_3)}{p(X_3)} \\ &= \frac{p(X_1)p(X_3)p(X_1|X_3)p(X_2|X_3)}{p(X_1)p(X_3)} \\ &= p(X_1|X_3)p(X_2|X_3). \end{aligned}$$

So we see that also in this case X_1 and X_2 are independent given X_3.
It is standard terminology to say that X_3 is *head-to-tail/tail-to-head* with respect to this path. Such a path does “connect” X_1 to X_2 but if we condition on X_3 then again this path is “blocked,” and the two variables become independent.

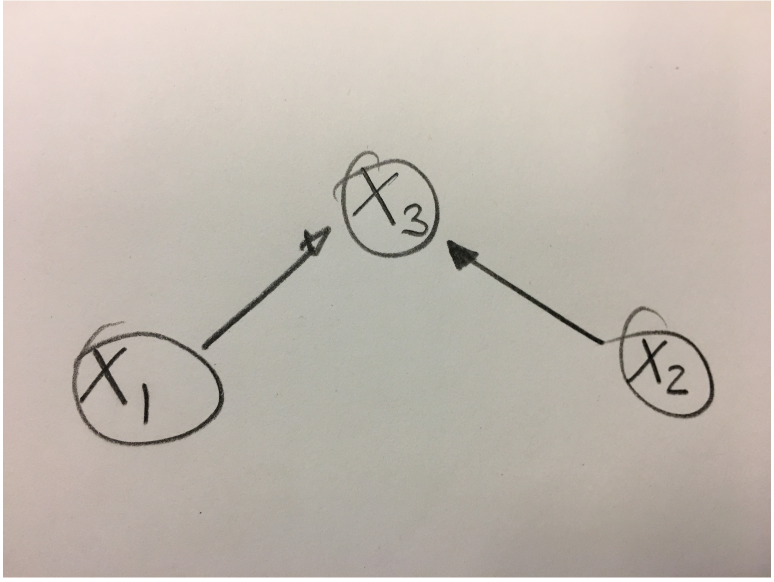

Figure 5: Third example: X_3 is *head-to-head* with respect to the path from X_1 to X_2. Conditioning *creates* dependence.

Let us now look at our final example. It is shown in Figure 5. This corresponds to the factorization

$$p(X_1, X_2, X_3) = p(X_1)p(X_2)p(X_3|X_1, X_2).$$

Marginalizing out X_3, we see that in this case X_1 and X_2 are independent. But if we condition on X_3 then generically we create a dependence. So, contrary to the previous two cases, a *head-to-head* path *creates* dependence if we condition on X_3.

The last example (and the fact that dependence is created) is at the core of a phenomen that is called *explaining away*. Assume that all the random variables are binary. Let X_1 be 1 mean that you are super smart (and $X_1 = 0$ mean that you are not so super smart :-)), $X_2 = 1$ mean that the final exam is super easy, and $X_3 = 1$ that you got a 6 in the final exam. A priori X_1 and X_2 might be independent. But if you

learn that $X_3 = 1$ then we create dependence: In particular, if there is now also evidence that $X_2 = 1$ this will change the probability that $X_1 = 1$ since $X_2 = 1$ already "explains away" why you passed the exam with a top grade.
Note that the same reasoning that showed us that conditioning on X_3 can create dependence applies also if X_3 has any descendents that are in Z.

D-Separation and Conditional Independence

Let us now state a general simple graphical criterion by which we can decide on (conditional) independence. We will see the same elemens that we encountered in our previous examples reappear.

Lemma. *The (set of) random variable(s) X is conditionally independent of the (set of) random variable(s) Y conditioned on the (set of random) variable(s) Z if X and Y are D-separated by Z. If X and Y are* not *D-separated by Z then we can construct a distribution that is compatibe with this Bayes' net so that X and Y are* not *conditionally independent given Z.*

Definition 0.1 (D-Separation)**.** *We say that X and Y are D-separated by Z (all of them can be sets of random variables) iff every path from any element of X to any element of Y is blocked by Z.*

Definition 0.2 (Blocked Path)**.** *We say that a path from a node X to a node Y is blocked by Z iff it contains a variable such that either*

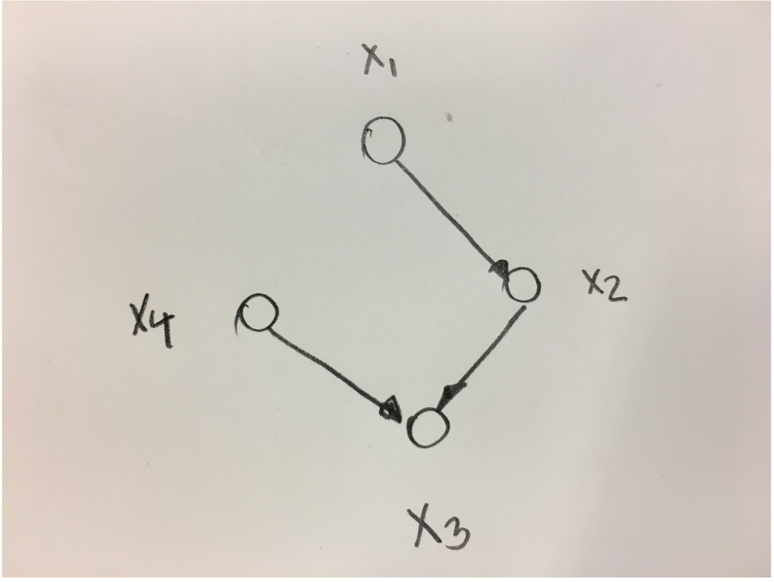

Figure 6: A Bayes net with four nodes.

1. *this variable is in Z and it is head-to-tail or tail-to-tail (as in our two first examples), or*

2. *the node is head-to-head and neither this node nor any of its descendants are in Z (like in our last example).*

Note: *Descendant* means child, or child of child, or ...
Consider the example shown in Figure 6.

- Is X_1 independent of X_3 given X_2? The answer is *yes*. There is only one path from X_1 to X_3. It goes through X_2 and X_2 is head-to-tails wrt this path. Therefore the only path is *blocked* by X_2 according to criterion 1 above.

- Is X_3 independent of X_1 given X_2? The answer is again *yes*. The notion of independence as well as our criteria above are symmetric.

- Is X_4 independent of X_1 given X_2? The answer is again *yes*. There is only one path from X_1 to X_4? It goes through X_2 and X_2 is head-to-tails wrt this path. Therefore the only path is *blocked* by X_2 according to criterion 1 above.

- Is X_4 independent of X_1 given X_3? The answer is *no*. Neither of the above two criteria apply.

- Is X_4 independent of X_1 given X_3 and X_2. The answer is *yes*.

- Is X_4 independent of X_1 given the empty set? The answer is *yes*. The only path between them is blocked at X_3 which is head-to-head, and neither X_3 nor any of its descendantas (it has none) belong to Z, which is the empty set.

Let us look at one more example. It is shown in Figure 7. We will not work the answers here, but it might be a good additional exercise to work out the answer yourself.

- Is X_1 independent of X_2 given X_3?

- Is X_1 independent of X_2 given X_5?

One direction of the Lemma is very easy to see. If some path from X to Y is not blocked by Z we need to show that we can construct a distribution that is compatible with this graphical model where X and Y are not independent given Z. The basic idea is the following. Take any path that is not blocked. We can fix any random variables outside this

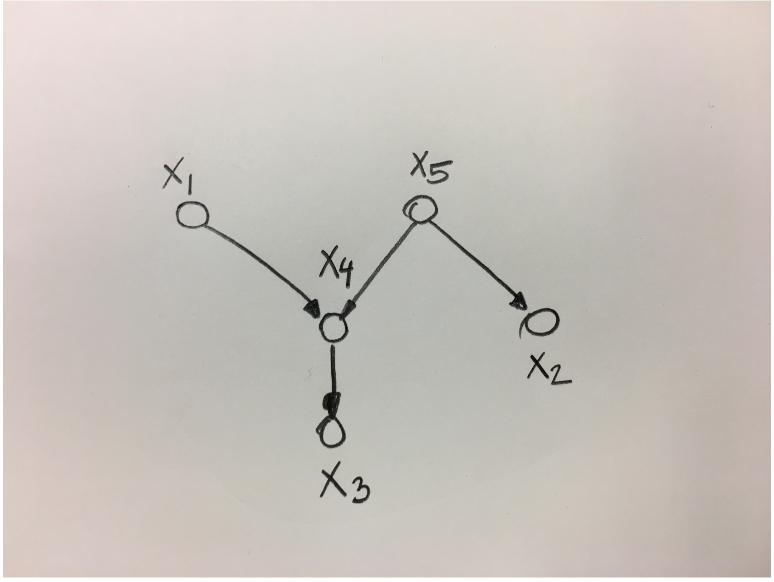

Figure 7: Another Bayes Net.

path to whatever we want. E.g., we could set them all to 0 with probability 1 and have them independent of all the variables in this path. This leaves only the variables on the path. So we only need to show that if we are given a graph that consists of a single path connecting X to Y that is not blocked by Z (which may or may not be on this path) then for this graph we can construct a distribution, compatible for this graph, where X and Y are conditionally dependent given Z. But this essentially boils down the the previous three examples.

The converse direction is more challenging. We skip the proof.

0.1 Causality

When looking at a Bayes' net, which is a directed graph, it might be tempting to interpret this as a causality relationship or an influence relationship. E.g., if you look at Figure 1, it

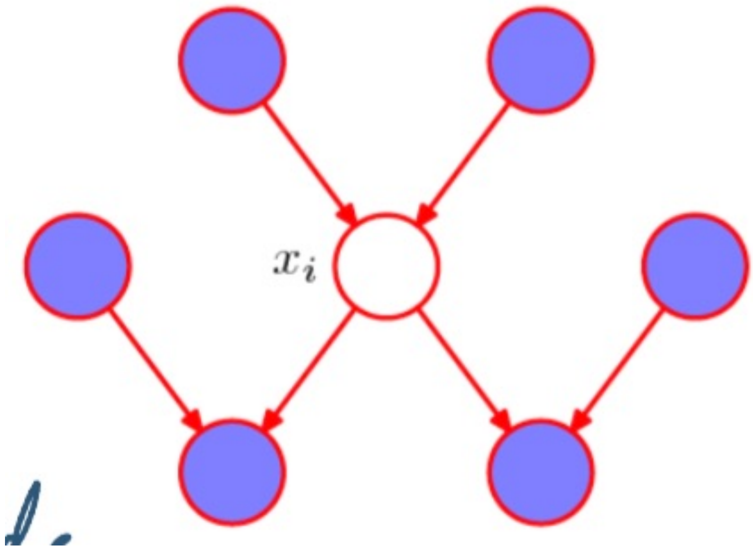

Figure 8: Markov blanket of X_i.

might appear that X is influenced by all other variables but does not influence any of them. But note that Figure 1 corresponds to the completely generic expansion in terms of the chain rule where we picked a particular ordering $1, 2, 3, 4$. We could have picked a different ordering and would then have arrived at a different graph.
All such a Bayes' Net can tell us is about conditional independence. I.e., the graph tells us that X and Y are D-separted given Z then we know that they are also independent given Z. But the converse is not necessarily true and we also cannot draw any conclusions regarding causality.

Markov Blanket

Given a node X_i we can ask if there is a kind of *minimal* set so that every random variable outside this set is conditionally independent of X_i. This is what is typically called the *Markov blanket*. Figure 8 shows such a blanket.

Definition 0.3 (Markov Blanket). *The Markov blanket of a node X_i is the set of parents, children, and co-parents of the node X_i. Here, by* co-parent, *we mean other parents of the children of X_i.*

It is a nice exercise to show that indeed any other node X_j which is not in the Markov blanket of X_i is conditionally independent of X_i given its Markov blanket by showing that X_j is D-separated from X_i by the Markov blanket.
Let us look at one of the cases. Fix X_i and let Z be its Markov blanket. Let X_j, $j \neq i$, and $X_j \notin Z$. Consider a path from X_i to X_j and assume that this path goes through a child of X_i (every path from X_i to X_j has to go through either a child or a parent of X_i). Let this child be Y. Note that Y must either be head-to-head or head-to-tail with respect to this path since the edge from X_i to Y is directed from X_i to Y (by assumption Y is a child of X_i). If the path is head-to-tail we are done since then this path is blocked by Y. But if Y is head-to-head then it must be true that the path also contains a co-parent of Y, call it U. Now this co-parent must either be tail-to-tail or tail-to-head with respect to this path. In both cases this co-parent blocks this path. The second case can be dealt with in a similar manner.

Sampling and Marginals

So far we have discussed how we can recognize independence relationship if we are given a Bayes net.
Perhaps even more important is the ability to compute marginals given a Bayes net or to be able to sample given a Bayes net.

These two tasks are related.
To see this, assume at first that we can sample efficiently given a Bayes net. To simplify things even further, assume that all variables are binary, i.e., $X_i \in \{0, 1\}$. We could then generate many independent samples $\{X_n\}_{n=1}^N = \{(X_{1n}, \cdots, X_{Dn})\}_{n=1}^N$. In order to estimate the marginal for X_i, i.e., in order to estimate $\mathbb{E}[X_i]$, we can then compute the corresponding empirical quantity $\frac{1}{N}\sum_{n=1}^N X_{in}$ and we know that this will converge to the true mean when we increase N.
Conversely, assume that we can efficiently compute marginals of any Bayes net. We want to sample from the joint distribution. We can then compute the marginal of the net with respect to X_1 lets say, and then flip a coin according to this marginal. We now have reduced the problem to generating a sample from the Bayes net where X_1 is already known and we can recurse.
The problem is that in general neither sampling nor computing marginals can be done efficiently except for special cases. If you look at (1) you see that in order to generate a sample by deciding on one variable at a time (taking the previous choices into account) then in order to sample X_i given the previous realizations of $X_1, \cdots, X_{i-1}$ we would need a table that has size (in the binary case) 2^{i-1} (this table contains the conditional probabilities). In other words, at least the storage complexity is exponential in the size of the net.
More generally, the storage requirement will be exponential in the largest number of parents any node in the network has.

Of course, if the Bayes net is a chain, then this task is very easy! In our next lecture we will talk about factor graphs. This is another graphical model. And we will ask how complex it is to compute marginals. This will lead us to the sum-product algorithm which is exact in the case where the factor graph is a tree and often gives a useful approximation even in the case where we have cycles.