Lecture topics:

- Bayesian networks

Bayesian networks

Bayesian networks are useful for representing and using probabilistic information. There are two parts to any Bayesian network model: 1) directed graph over the variables and 2) the associated probability distribution. The graph represents qualitative information about the random variables (conditional independence properties), while the associated probability distribution, consistent with such properties, provides a quantitative description of how the variables relate to each other. If we already have the distribution, why consider the graph? The graph structure serves two important functions. First, it explicates the properties about the underlying distribution that would be otherwise hard to extract from a given distribution. It is therefore useful to maintain the consistency between the graph and the distribution. The graph structure can also be learned from available data, i.e., we can explicitly learn qualitative properties from data. Second, since the graph pertains to independence properties about the random variables, it is very useful for understanding how we can use the probability model efficiently to evaluate various marginal and conditional properties. This is exactly why we were able to carry out efficient computations in HMMs. The forward-backward algorithms relied on simple Markov properties which are independence properties, and these are generalized in Bayesian networks. We can make use of independence properties whenever they are explicit in the model (graph).

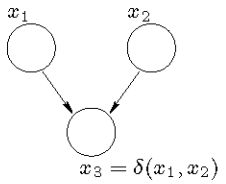

Figure 1: A simple Bayesian network over two independent coin flips x_1 and x_2 and a variable x_3 checking whether the resulting values are the same. All the variables are binary.

Let's start with a simple example Bayesian network over three binary variables illustrated in Figure 1. We imagine that two people are flipping coins independently from each other. The resulting values of their unbiased coin flips are stored in binary (0/1) variables x_1 and x_2. Another person checks whether the coin flips resulted in the same value and the

outcome of the comparison is a binary (0/1) variable $x_3 = \delta(x_1, x_2)$. Based on the problem description we can easily write down a joint probability distribution over the three variables

$$P(x_1, x_2, x_3) = P(x_1)P(x_2)P(x_3|x_1, x_2) \quad (1)$$

where $P(x_1) = 0.5$, $x_1 \in \{0, 1\}$, $P(x_2) = 0.5$, $x_2 \in \{0, 1\}$, and $P(x_3 = 1|x_1, x_2) = 1$ if $x_1 = x_2$ and zero otherwise.

We could have read the structure of the joint distribution from the graph as well. We need a bit of terminology to do so. In the graph, x_1 is a *parent* of x_3 since there's a directed edge from x_1 to x_3 (the value of x_3 depends on x_1). Analogously, we can say that x_3 is a *child* of x_1. Now, x_2 is also a parent of x_3 so that the value of x_3 depends on both x_1 and x_2. We will discuss later what the graph means more formally. For now, we just note that Bayesian networks always define acyclic graphs (no directed cycles) and represent how values of the variables depend on their parents. As a result, any joint distribution consistent with the graph, i.e., any distribution we could imagine associating with the graph, has to be able to be written as a product of conditional probabilities of each variable given its parents. If a variable has no parents (as is the case with x_1) then we just write $P(x_1)$. Eq.(1) is exactly a product of conditional probabilities of variables given their parents.

Marginal independence and induced dependence

Let's analyze the properties of the simple model a bit. For example, what is the marginal probability over x_1 and x_2? This is obtained from the joint simply by summing over the values of x_3

$$P(x_1, x_2) = \sum_{x_3} P(x_1)P(x_2)P(x_3|x_1, x_2) = P(x_1)P(x_2) \sum_{x_3} P(x_3|x_1, x_2) = P(x_1)P(x_2) \quad (2)$$

Thus x_1 and x_2 are *marginally independent* of each other. In other words, if we don't know the value of x_3 then there's nothing that ties the coin flips together (they were, after all, flipped independently in the description). This is also a property we could have extracted directly from the graph. We will provide shortly a formal way of deriving this type of independence properties from the Bayesian network.

Another typical property of probabilistic models is *induced dependence*. Suppose now that the coins x_1 and x_2 were flipped independently but we don't know their outcomes. All we know is the value of x_3, i.e., whether the outcomes where identical or not (say they were identical). What do we know about x_1 and x_2 in this case? We know that either $x_1 = x_1 = 0$ or $x_1 = x_2 = 1$. So their values are clearly *dependent*. The dependence was *induced by additional knowledge*, in this case the value of x_3. This is again a property we

could have read off directly from the graph (explained below). Note that the dependence pertains to our beliefs about the values of x_1 and x_2. The coins were physically flipped independently of each other and our knowledge of the value of x_3 doesn't change this. However, the value of x_3 narrows down the set of possible outcomes of the two coin flips for this *particular sample of* x_1 *and* x_2.

Both marginal independence and induced dependence are typical properties of realistic models. Consider, for example, a factorial Hidden Markov Model in Figure 2c). In this model you have two marginally independent Markov models that conspire to generate the observed output. In other words, the two Markov models are tied only through observations (induced dependence). To sample values for the variables in the model, we would be sampling from the two Markov models independently and just using the two states at each time point to sample a value for the output variables. The joint distribution over the variables for the model in Figure 2c) is again obtained by writing a product of conditional probabilities of each variable given its parents:

$$P(x_1')P(x_1)P(y_1|x_1', x_1)P(x_2'|x_1')P(x_2|x_1)P(y_2|x_2, x_2')P(x_3'|x_2')P(x_3|x_2)P(y_3|x_3, x_3) \quad (3)$$

where, e.g., $P(y_1|x_1', x_1)$ could be defined as $N(y; \mu(x_1') + \mu(x_1), \sigma^2)$. Such a model could, for example, capture how two independent subprocesses in speech production generate the observed acoustic signal, model two speakers observed through a common microphone, or with a different output model, capture how haplotypes generate observed genotypes. Given the model and say an observed speech signal, we would be interested in inferring likely sequences of states for the subprocesses.

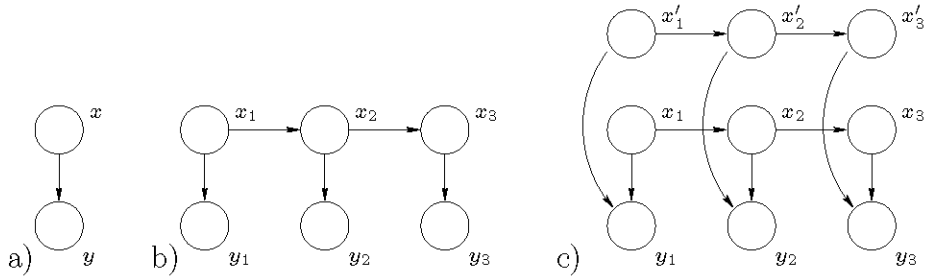

Figure 2: Different models represented as Bayesian networks: a) mixture model, b) HMM, c) factorial HMM.

Explaining away

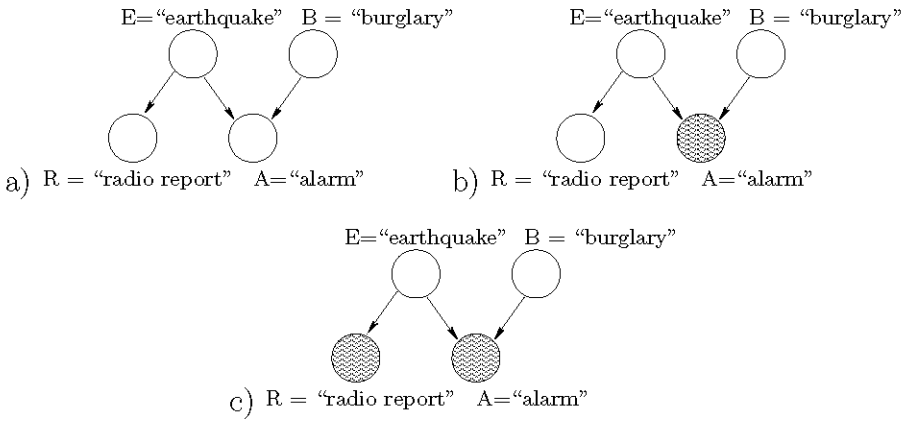

Figure 3: Network structure exhibiting explaining away. a) basic model, b) alarm went off, c) we have also heard a radio report about an earthquake

Another typical phenomenon that probabilistic models can capture is *explaining away*. Consider the following typical example (Pearl 1988) in Figure 3. We have four variables A, B, E, and R capturing possible causes for why a burglary alarm went off. All the variables are binary (0/1) and, for example, $A = 1$ means that the alarm went off (Figure 3b). Shaded nodes indicate that we know something about the values of these variables. In our example here all the observed values are one (property is true). We assume that earthquakes ($E = 1$) and burglaries ($B = 1$) are equally unlikely events $P(E = 1) = P(B = 1) \approx 0$. Alarm is likely to go off only if either $E = 1$ or $B = 1$ or both. Both events are equally likely to trigger the alarm so that $P(A = 1|E, B) \approx A$ or B. An earthquate ($E = 1$) is likely to be followed by a radio report ($R = 1$), $P(R = 1|E = 1) \approx 1$, and we assume that the report never occurs unless an earthquake actually took place: $P(R = 1|E = 0) = 0$.

What do we believe about the values of the variables if we only observe that the alarm went off ($A = 1$)? At least one of the potential causes $E = 1$ or $B = 1$ should have occured. However, since both are unlikely to occur by themselves, we are basically left with either $E = 1$ or $B = 1$ but (most likely) not both. We therefore have two alternative or competing explanations for the observation and both explanations are equally likely. If we know hear, in addition, that there was a radio report about an earthquake, we believe that $E = 1$. This makes $B = 1$ unnecessary for explaining the alarm. In other words, the additional observation about the radio report *explained away* the evidence for $B = 1$. Thus, $P(E = 1|A = 1, R = 1) \approx 1$ whereas $P(B = 1|A = 1, R = 1) \approx 0$.

Note that we have implicitly captured in our calculation here that R and B are *dependent*

given $A = 1$. If they were not, we would not be able to learn anything about the value of B as a result of also observing $R = 1$. Here the effect is drastic and the variables are strongly dependent. We could have, again, derived this property from the graph.

Bayesian networks and conditional independence

We have claimed in several occasions that we could have derived useful properties about the probability model directly from the graph. How is this done exactly? Since the graph encodes independence properties about the variables, we have to define a criterion for extracting independence properties between the variables directly from the graph. For Bayesian networks (acyclic graphs) this is given by so called *D-separation criterion.*

As an example, consider a slightly extended version of the previous model in Figure 4a, where we have added a binary variable L (whether we "leave work" as a result of hearing/learning about the alarm). We will define a procedure for answering questions such as: are R and B independent given A?

The general procedure involves three graph transformation steps that we will illustrate in relation to the graph in Figure 4a.

1. Construct *ancestral graph* of the variables of interest. The variables we care about here are R, B, and A. The ancestral graph includes these variables as well as all the variables (ancestors) you can get to by starting from one of these variables and following the arrows in the reverse direction (their parents, their parents' parents, and so on). The ancestral graph in our case is given in Figure 4b.

   The motivation for this step is that unobserved effects of random variables cannot lead to dependence and can be therefore removed.

2. *Moralize* the resulting ancestral graph. This operation simply adds an *undirected edge* between any two variables in the ancestral graph that have a common child ("marry the parents"). In case of multiple parents, they are connected pairwise, i.e., by adding an edge between any two parents. See Figure 4c.

   Moralization is needed to take into account induced dependences discussed earlier.

3. Change all the direct edges into undirected edges. This gives the resulting *undirected graph* in Figure 4d.

We can now read off the answer to the original question from the resulting undirected graph. R and B are independent given A (they are *D-separated given* A) if they are separated by A in the undirected graph. In other words, if they become disconnected in the undirected

graph by removing the conditioning variable A and its associated edges. They clearly remain connected in our example and thus, from the point of view of the Bayesian network model, would have to be assumed to be dependent.

Let's go back to the previous examples to make sure we can read off the properties we claimed from the graphs. For example, if we are interested in asking whether x_1 and x_2 are marginally independent (i.e., given nothing) in the model in Figure 1, we would create the graph transformations shown in Figure 5. The nodes are clearly separated. Similarly, to establish that x_1 and x_2 become dependent with the observation of x_3, we would ask whether x_1 and x_2 are independent given x_3 and get the transformations in Figure 6. The nodes are not separated by x_3 and therefore not independent.

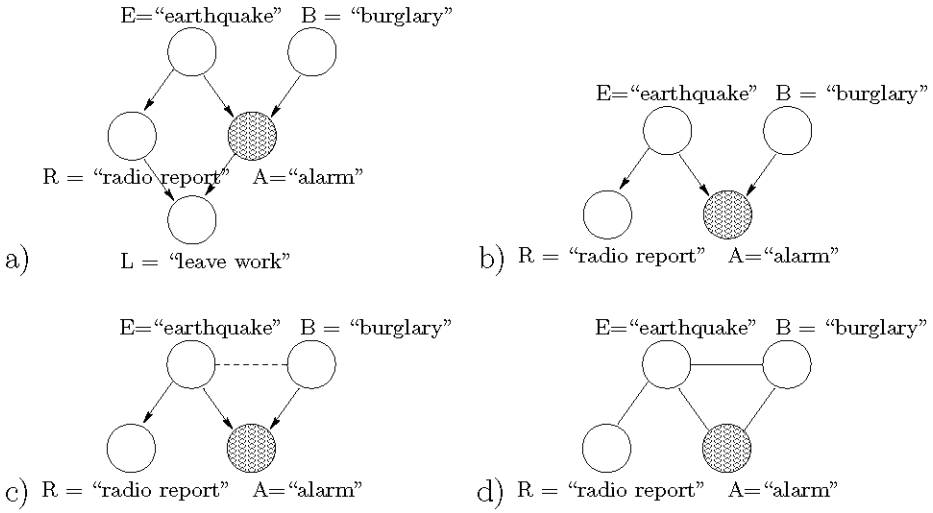

Figure 4: a) Burglary model, extended, b) ancestral graph of R, B, and A, c) moralized ancestral graph, d) resulting undirected graph.

Graph and the probability distribution

The graph and the independence properties we can derive from it are useful to us only if the probability distribution we associate with the graph is consistent with the graph. By consistency we meant that all the independence properties we can derive from the graph should hold for the associated distribution. In other words, if the graph is an explicit representation of such properties, then clearly whatever we can infer from it, should be true. There are actually a large number of possible independence properties that we can derive from any typical graph, even in the context of HMMs. How is it that we can ever hope to find and deal with distributions that are consistent with all such properties? While

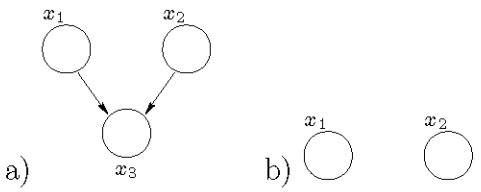

Figure 5: a) Bayesian network model, b) ancestral graph of x_1 and x_2, already moralized and undirected.

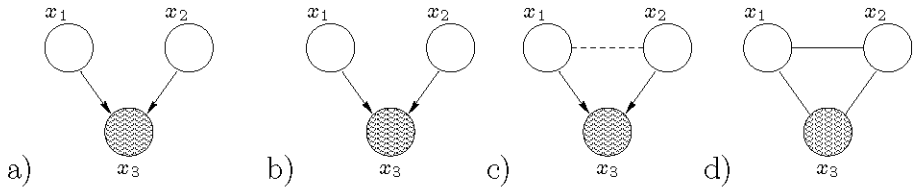

Figure 6: a) Bayesian network model, b) ancestral graph of x_1 and x_2 given x_3, c) moralized ancestral graph, d) resulting undirected graph.

the task is hard, the answer is simple. In fact, given an acyclic graph G over d variables, the most general form of the joint distribution consistent with *all* the properties we can derive from the graph is given by

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{d} P(x_i | x_{pa_i}) \tag{4}$$

where x_{pa_i} refers to the set of variables that are the parents of variable x_i (e.g., $x_{pa_3} = \{x_1, x_2\}$ for x_3 in the above models). So, we can just read off the answer from the graph: look at each variable and include a term in the joint distribution of that variable given its parents (those that directly influence it).

Note that some distributions may satisfy more independence properties that are represented in the graph. For example, a distribution where all the variables are independent of each other is consistent with every acyclic graph. It clearly satisfies all the possible independence properties (edges in the graph only indicate possible dependences; they may actually be

weak or non-existent). We typically would use a graph representation that tries to capture most if not all of the independence properties that hold for the associated distribution. Not all independence properties can be captured (are representable) by our D-separation criterion.