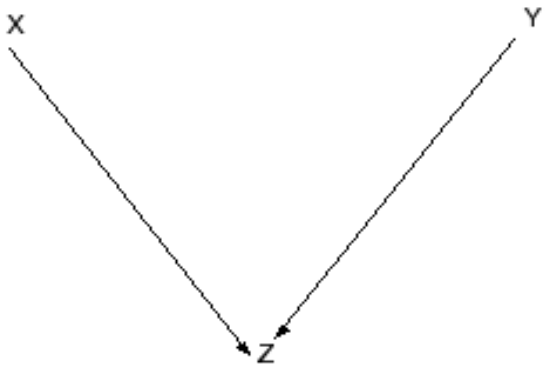
An Introduction to Probabilistic Graphical Models

Reading:

- Chapters 17 and 18 in Wasserman.

Directed Graphs

We wish to identify simple structure in large and complex probabilistic models arising e.g. in sensor networks. Graphical models are a suitable tool for this purpose.



Definition 1. *A directed graph consists of nodes (or vertices)* $X, Y, \ldots$ *and arrows (or edges) connecting some of the nodes. More formally, we define a set of vertices* $\mathcal{V}$ *and an edge set* $\mathcal{E}$ *of ordered pairs of vertices.*

Definition 2. *Consider random variables* $X, Y,$ *and* Z*.* X *and* Y *are conditionally independent given* Z*, written*

$$X \perp\!\!\!\perp Y \,|\, Z$$

if

$$f_{X,Y\,|\,Z}(x, y \,|\, z) = f_{X\,|\,Z}(x \,|\, z)\, f_{Y\,|\,Z}(y \,|\, z)$$

for all $x, y,$ *and* z*.*
In words: Knowing Z *renders* Y *irrelevant for predicting* X*. Knowing* Z *renders* X *irrelevant for predicting* Y*.*

Lemma. *Clearly,*

$$X \perp\!\!\!\perp Y \,|\, Z$$

if and only if

$$f_{X \,|\, Y,Z}(x \,|\, y, z) = f_{X \,|\, Z}(x \,|\, z).$$

Theorem 1. *The following implications hold:*

$$X \perp\!\!\!\perp Y \,|\, Z \quad \Longrightarrow Y \perp\!\!\!\perp X \,|\, Z \tag{1}$$

$$Y \perp\!\!\!\perp X \,|\, Z \quad \Longrightarrow Y \perp\!\!\!\perp h(X) \,|\, Z \tag{2}$$

$$Y \perp\!\!\!\perp X \,|\, Z \quad \Longrightarrow Y \perp\!\!\!\perp X \,|\, \{Z, h(X)\} \tag{3}$$

$$Y \perp\!\!\!\perp X \,|\, Z \quad \textit{and} \quad W \perp\!\!\!\perp X \,|\, \{Y, Z\} \Longrightarrow \{Y, W\} \perp\!\!\!\perp X \,|\, Z \tag{4}$$

$$Y \perp\!\!\!\perp X \,|\, Z \quad \textit{and} \quad Z \perp\!\!\!\perp X \,|\, Y \Longrightarrow \{Y, Z\} \perp\!\!\!\perp X. \tag{5}$$

[Property (5) requires that all its events have positive probability.]

We show (2) assuming the discrete-distribution case, for simplicity. We know

$$p_{X,Y \,|\, Z}(x, y \,|\, z) = p_{X \,|\, Z}(x \,|\, z) \cdot p_{Y \,|\, Z}(y \,|\, z)$$

and, therefore, for $U = h(X)$,

$$\begin{aligned} p_{U,Y\,|\,Z}(u, y\,|\,z) &= \sum_{\xi:\, h(\xi)=u} p_{X,Y\,|\,Z}(\xi, y\,|\,z) \\ &= \underbrace{\sum_{\xi:\, h(\xi)=u} p_{X\,|\,Z}(\xi\,|\,z)}_{p_{U\,|\,Z}(u\,|\,z)} \cdot p_{Y\,|\,Z}(y\,|\,z) \end{aligned}$$

i.e.

$$Y \perp\!\!\!\perp \underbrace{U}_{h(X)} \,|\, Z.$$

Proof of (3):

$$Y \perp\!\!\!\perp X \,|\, Z$$

means

$$f_{Y\,|\,X,Z}(y\,|\,x, z) = f_{Y\,|\,Z}(y\,|\,z)$$

which further implies

$$f_{Y\,|\,Z}(y\,|\,z) = f_{Y\,|\,X,Z}(y\,|\,x, z) = f_{Y\,|\,X,h(X),Z}(y\,|\,x, h(x), z)$$

i.e.

$$Y \perp\!\!\!\perp X \,|\, \{h(X), Z\}.$$

Possibly a more natural statement than (3) is

$$Y \perp\!\!\!\perp X \,|\, Z \quad \Longrightarrow \quad Y \perp\!\!\!\perp \{X, h(X)\} \,|\, Z$$

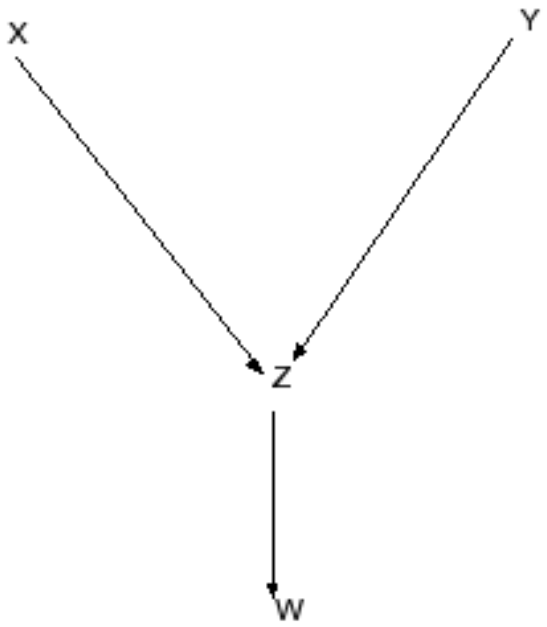
which is equivalent to (3).

Definition 3. *If an arrow (pointing in either direction) connects nodes X and Y, we call these nodes adjacent.*

Definition 4. *If an arrow points from X to Y, we say that X is a parent of Y and Y is a child of X.*

Definition 5. *A set of arrows beginning at X and ending at Y is called a directed path between X and Y.*

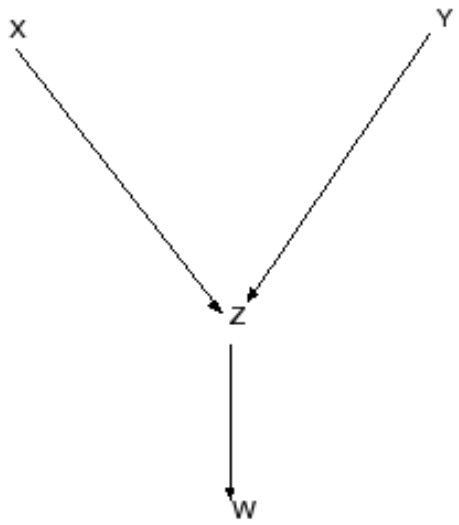
Example:



In this example (in the above figure), we have a directed path from X to W and a directed path from Y to W.

Definition 6. *A sequence of adjacent vertices starting at X and ending at Y without reference to the direction of the arrows is an undirected path.*

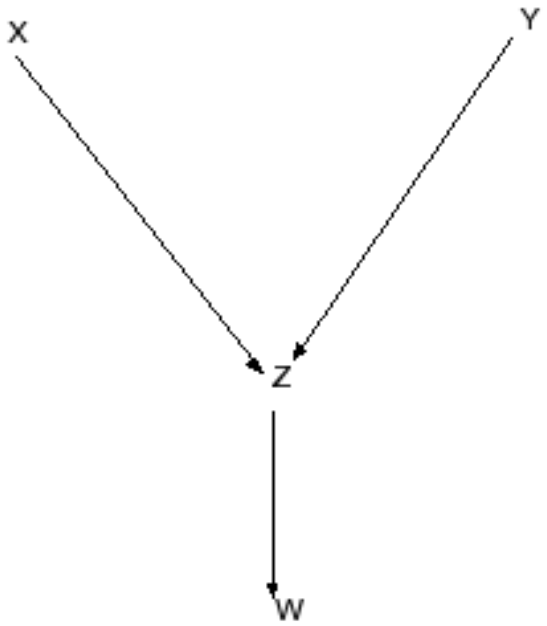
Definition 7. *If there is a directed path from X to Y (or if $X = Y$), we say that X is an ancestor of Y and that Y is a descendant of X.*



- X and Z are adjacent,
- X and Y are not adjacent,
- X is a parent of Z,
- X is an ancestor of W,
- Z is a child of X,
- W is a descendant of X,
- there is a directed path from X to W,
- there is an undirected path from X to W
- there is an undirected path from X to Y.

Definition 8. *A sequence of vertices constituting an undirected path X to Y has a collider at Z if there are two arrows along the path pointing to Z.*

Definition 9. *When vertices with arrows pointing into a collider at Z are not adjacent, then we say that the collider is unshielded.*



An undirected path X to Y has a collider at Z. Z is an *unshielded collider* on the undirected path $X - Z - Y$.

On the undirected path $X - Z - W$, Z is *not a collider!*

Definition 10. *A directed path that starts and ends at the same vertex is called a cycle. A directed graph is called acyclic if it has no cycles.*

Abbreviation: DAG $\equiv$ directed acyclic graph.

Denote by $\mathcal{G}$ a DAG. From now on, as far as directed graphs are concerned, we only deal with DAGs.

Definition 11. *Consider a DAG $\mathcal{G}$ with vertices*

$$\boldsymbol{X} = [X_1, X_2, \ldots, X_K]^T$$

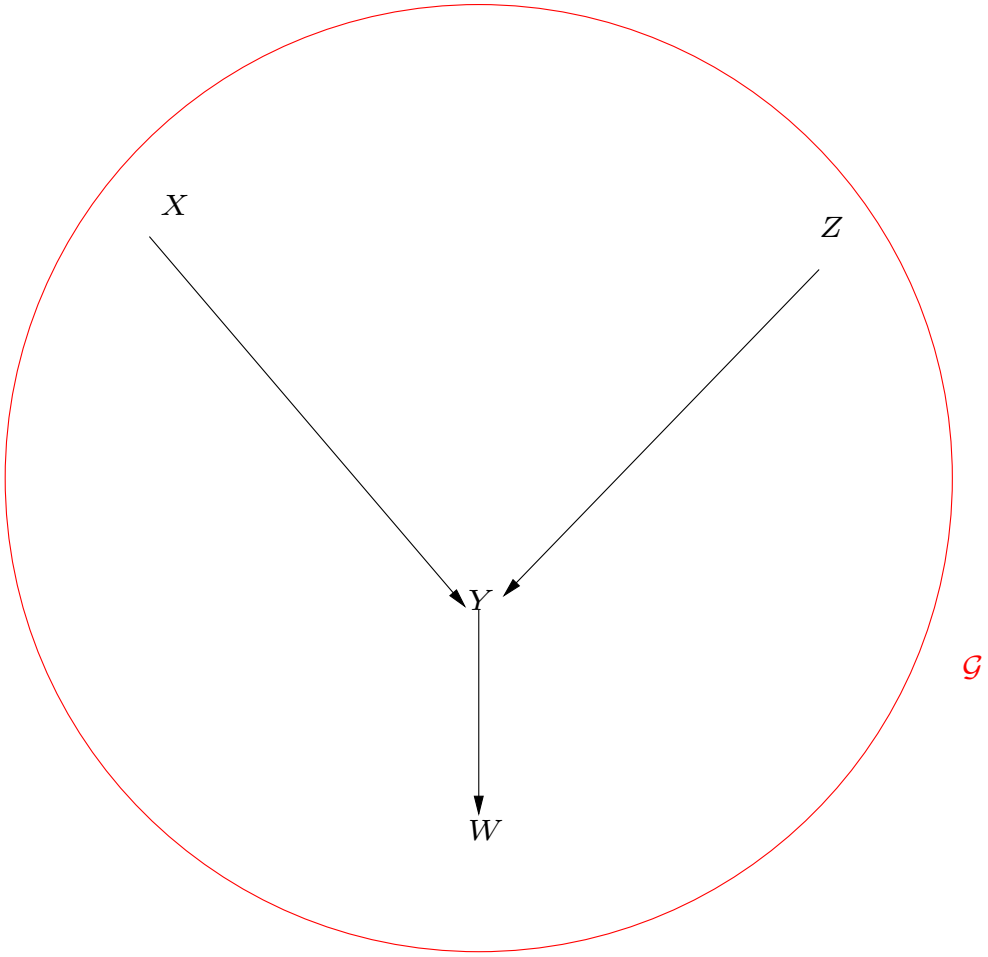
where "T" denotes a transpose. Then, a distribution F for $\boldsymbol{X}$ is Markov to $\mathcal{G}$ *or $\mathcal{G}$* represents F *if and only if*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{K} p_{X_i \,|\, X_{\mathrm{pa}_i}}(x_i \,|\, x_{\mathrm{pa}_i})$$

where $\mathrm{pa}_i \equiv$ set of parents of node i in $\mathcal{G}$. The set of distributions F that are represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

Notational Comments: Here, we interchangeably denote a node either by its index i or by its random variable X_i. Also, we adopt the following notation: $X_{\cdot} = \{X_j \,|\, j \in \cdot\}$. For example, $X_{\mathrm{pa}_i} = \{X_j \,|\, j \in \mathrm{pa}_i\}$.

Example:



What does it mean to say that $\mathcal{G}$ in the above figure *represents* a joint distribution for $[X, Y, Z, W]^T$?

Answer:

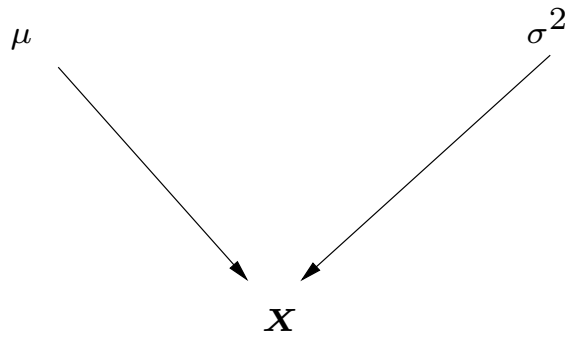
$$f_{X,Y,Z,W}(x, y, z, w) = p_X(x) \cdot p_Z(z) \cdot p_{Y \mid X,Z}(y \mid x, z) \cdot p_{W \mid Y}(w \mid y).$$

Example: Suppose that our measurement vector $\boldsymbol{X} = \boldsymbol{x}$ given μ, σ^2 follows

$$f_{\boldsymbol{X} \,|\, \mu, \Sigma^2}(\boldsymbol{x} \,|\, \mu, \sigma^2)$$

and that we choose independent priors for μ and σ^2:

$$f_{\mu, \Sigma^2}(\mu, \sigma^2) = f_{\mu}(\mu)\, f_{\Sigma^2}(\sigma^2).$$



The joint pdf of $\boldsymbol{X}, \mu$, and Σ^2 is

$$f_{\boldsymbol{X}, \mu, \Sigma^2}(\boldsymbol{x}, \mu, \sigma^2) = f_{\mu}(\mu)\, f_{\Sigma^2}(\sigma^2)\, f_{\boldsymbol{X} \,|\, \mu, \Sigma^2}(\boldsymbol{x} \,|\, \mu, \sigma^2).$$

In general, if a DAG $\mathcal{G}$ represents F, we can say that

$$f_{X_i \,|\, (\boldsymbol{x} \backslash X_i)}(x_i \,|\, (\boldsymbol{x} \backslash x_i)) \propto \text{terms in } f(\boldsymbol{x}) \text{ containing } x_i$$

$$\propto \overbrace{f_{X_i \,|\, X_{\mathrm{pa}_i}}(x_i \,|\, x_{\mathrm{pa}_i})}^{\text{"prior"}} \cdot \underbrace{\prod_{j,\, x_i \in x_{\mathrm{pa}_j}} f_{X_j \,|\, X_{\mathrm{pa}_j}}(x_j \,|\, x_{\mathrm{pa}_j})}_{\text{"likelihood"}}$$

where $(\boldsymbol{x} \backslash x_i)$ is the collection of all vertices except x_i (read "$\boldsymbol{x}$ remove x_i"). The first term is the pdf (pmf) of x_i given its parents and the second is the product of pdfs (pmfs) in which x_i is a parent.
Hence, the nodes that are involved in the above full conditional pdf (pmf) are: X_i, its parents, children, and co-parents, where co-parents are defined as nodes that share a child (at least one, could be more).

Theorem 2. *A distribution is represented by a DAG $\mathcal{G}$ if and only if, for every X_i,*

$$X_i \perp\!\!\!\perp \widetilde{X}_i \,|\, X_{\mathrm{pa}_i} \quad \textit{Markov condition}$$

where $\widetilde{X}_i$ stands for all other entries of $\boldsymbol{X}$ except parents and descendants of X_i.

Proof. (A Rough Proof.) Adopt a topological ordering of the

nodes such that $\mathrm{pa}_i \subset \{1, 2, \ldots, i-1\}$, which we can always do on a DAG. Here, we focus on

$$\widetilde{X}_i = \{X_k \,|\, k \in \{1, 2, \ldots, i-1\} \backslash \mathrm{pa}_i\}.$$

(This $\widetilde{X}_i$ is not necessarily the same as the $\widetilde{X}_i$ in the statement of the theorem, but we can use this definition of $\widetilde{X}_i$ without loss of generality, due to the fact that we can always find node ordering so that this $\widetilde{X}_i$ is the same as that in the statement of Theorem 2).

Suppose first that $\mathcal{G}$ represents F, implying

$$\begin{aligned} p(x_1, x_2, \ldots, x_i) &= \sum_{x_{i+1}, \ldots, x_K} p(x_1, x_2, \ldots, x_K) \\ &= p(x_1)\, p(x_2 \,|\, x_{\mathrm{pa}_2}) \ldots p(x_i \,|\, x_{\mathrm{pa}_i}). \end{aligned}$$

We wish to prove that $X_i \perp\!\!\!\perp \widetilde{X}_i \,|\, X_{\mathrm{pa}_i}$ or, equivalently, that

$$\underbrace{p(x_i \,|\, \widetilde{x}_i, x_{\mathrm{pa}_i})}_{p(x_i \,|\, x_1, \ldots, x_{i-1})} = p(x_i \,|\, x_{\mathrm{pa}_i}).$$

Clearly

$$p(x_1) = p(x_1 \,|\, x_{\mathrm{pa}_1}) = p(x_1 \,|\, \emptyset)$$

since X_1 has no parents, and, consequently,

$$p(x_1 \,|\, \widetilde{x}_1, x_{\mathrm{pa}_1}) = p(x_1 \,|\, x_{\mathrm{pa}_1}).$$

For X_1 and X_2, we have

$$p(x_1, x_2) \overset{\text{chain rule}}{=} p(x_2 \,|\, x_1) \cdot p(x_1) = p(x_2 \,|\, x_{\mathrm{pa}_2}) \cdot p(x_1)$$

implying

$$p(x_2 \,|\, x_1) = p(x_2 \,|\, x_{\mathrm{pa}_2}).$$

Assume now that

$$p(x_i \,|\, x_1, \ldots, x_{i-1}) = p(x_i \,|\, x_{\mathrm{pa}_i}).$$

for all $i \leq j$ and consider $j+1$:

$$p(x_1, \ldots, x_{j-1}, x_j, x_{j+1}) = p(x_1)\, p(x_2 \,|\, x_1) \cdots p(x_{j+1} \,|\, x_1, \ldots, x_j)$$

$$\overset{\text{ind. hyp.}}{=} p(x_1)\, p(x_2 \,|\, x_{\mathrm{pa}_2}) \cdots p(x_j \,|\, x_{\mathrm{pa}_j})\, p(x_{j+1} \,|\, x_1, \ldots, x_j)$$

$$\overset{\mathcal{G} \text{ represents } F}{=} p(x_1)\, p(x_2 \,|\, x_{\mathrm{pa}_2}) \cdots p(x_j \,|\, x_{\mathrm{pa}_j})\, p(x_{j+1} \,|\, x_{\mathrm{pa}_{j+1}})$$

which implies

$$p(x_{j+1} \,|\, x_1, \ldots, x_j) = p(x_{j+1} \,|\, x_{\mathrm{pa}_{j+1}})$$

thus completing the induction proof.

The other direction is easy: we start from

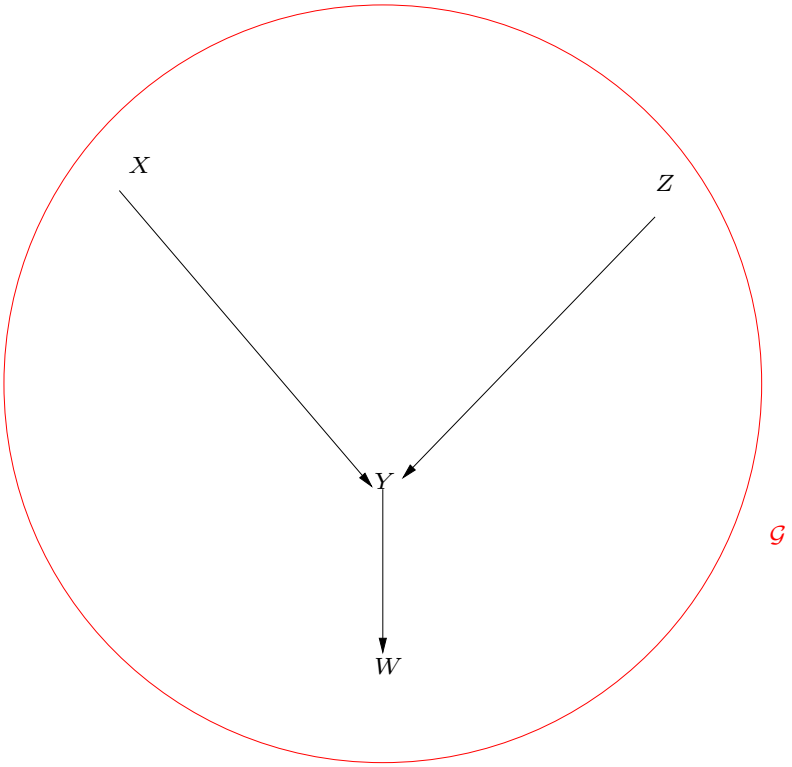
$$\underbrace{p(x_i \,|\, \widetilde{x}_i, x_{\mathrm{pa}_i})}_{p(x_i \,|\, x_1, \ldots, x_{i-1})} = p(x_i \,|\, x_{\mathrm{pa}_i})$$

and the chain rule immediately gives us

$$p(x_1, \ldots, x_K) = p(x_1) \underbrace{p(x_2 \,|\, x_1)}_{p(x_2 \,|\, x_{\mathrm{pa}_2})} \cdots \underbrace{p(x_K \,|\, x_1, \ldots, x_{K-1})}_{p(x_K \,|\, x_{\mathrm{pa}_K})}$$

which directly implies that $\mathcal{G}$ represents F. $\square$

(Back to) Example:



What about conditional independence between X and Z? X has no parents, so conditioning is on nothing. Hence, if $\mathcal{G}$ represents F, Theorem 2 tells us that

$$X \perp\!\!\!\perp Z.$$

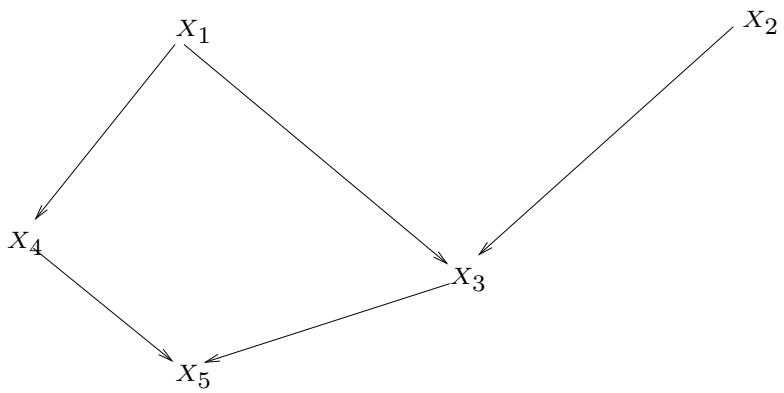
We can also conclude that

$$W \perp\!\!\!\perp \{X, Z\} \mid Y.$$

Here, the conditioning is on Y because Y is the parent of W.

Question: In addition to the results of Theorem 2, what other conditional independence relationships follow from the fact that F is represented by DAG $\mathcal{G}$?

Example: Suppose that this graph represents a probability distribution F:



What do we know about F that is represented by this graph? The Markov condition in Theorem 2 implies the following

(condition on parents, exclude parents and descendants):

$$
\begin{aligned}
X_1 &\perp\!\!\!\perp X_2 \qquad (X_1 \text{ and } X_2 \text{ have no parents}) \\
X_2 &\perp\!\!\!\perp \{X_1, X_4\} \\
X_3 &\perp\!\!\!\perp X_4 \mid \{X_1, X_2\} \\
X_4 &\perp\!\!\!\perp \{X_2, X_3\} \mid X_1 \\
X_5 &\perp\!\!\!\perp \{X_1, X_2\} \mid \{X_3, X_4\}.
\end{aligned}
$$

Furthermore, Theorem 2 tells us that the above relationships are *equivalent* to the Markov property in Definition 11. But, they do not exhaust all that can be said; for example,

$$X_2 \perp\!\!\!\perp \{X_4, X_5\} \mid \{X_1, X_3\}.$$

This is true, but does not immediately follow from Theorem 2.

To easily identify independence relationships beyond the definition of "$\mathcal{G}$ represents F" (Definition 11) or Theorem 2, we need new results.

Definition 12. *Let i and j be distinct vertices of a DAG and Q be a set of vertices not containing i or j. Then, X and Y are d-connected given Q if there is an undirected path $\mathcal{P}$ between i and j such that*

(i) *every collider in $\mathcal{P}$ has a descendant in Q and*

(ii) *no other vertex [besides possibly those mentioned in* (i)*] on* $\mathcal{P}$ *is in* Q*.*

If i *and* j *are not d-connected given* Q*, they are d-separated given* Q*.*

Abbreviation: d-separation $\equiv$ directed separation etc.

Definition 13. *If* A, B*, and* Q *are non-overlapping sets of vertices in a DAG and* A *and* B *are not empty, then*

$$A \text{ and } B \text{ are d-separated given } Q$$

if, for every $i \in A$ *and* $j \in B$*,* i *and* j *are d-separated given* Q*. If* A *and* B *are not d-separated given* Q*, they are d-connected given* Q*.*

Theorem 3. *Let* A, B*, and* C *be disjoint sets in a DAG representing* F*. Then*

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

if and only if A *and* B *are d-separated by* C*.*
(Recall that $X_\cdot = \{X_i \,|\, i \in \cdot\}$*.)*

Bayes-Ball Approach to Determining d-Separation Between Node Sets A and B

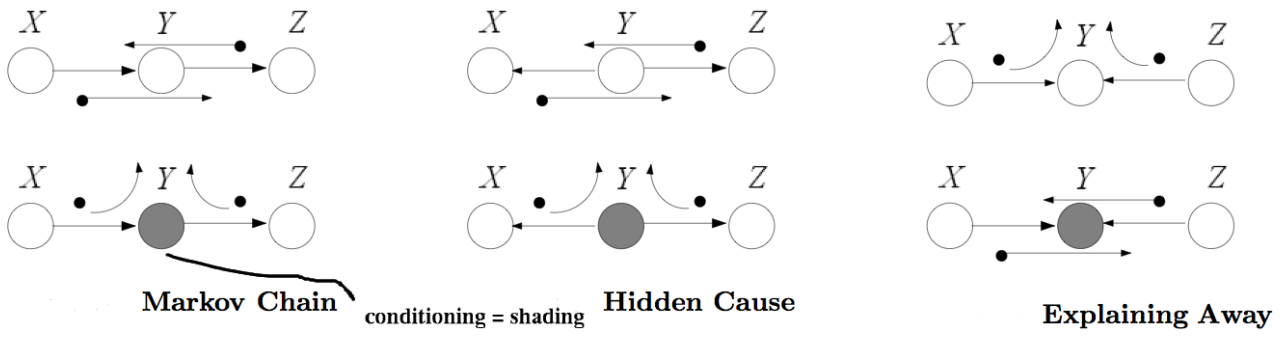
1. First, mark (e.g. shade) the nodes C that are conditioned on.

2. Start the ball within the nodes in set A and bounce it around the graph according to the conditional-independence rules stated below.

3. Finally, evaluate the results:

   - if the ball can reach a node in B, then A and B are d-connected,
   - if the ball cannot reach B, then the nodes in A and B are d-separated.

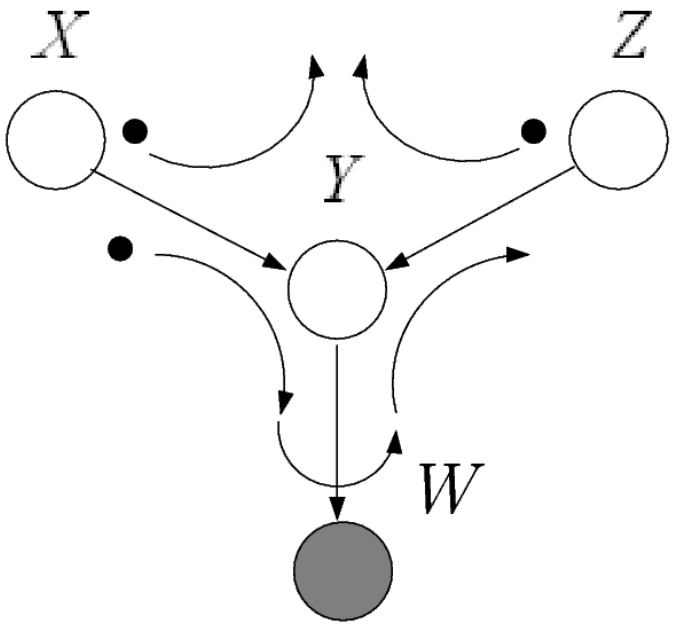
Bayes-ball rules. Here are the rules that govern the bouncing of the ball:



In words: When moving from X to Z (or Z to X) in the above canonical graphs,

- when Y is not a collider, the ball passes through Y if we *do not* condition on Y;
- when Y is not a collider, the ball bounces off of Y if we *condition* on Y;
- when X and Z collide at Y, the ball bounces off of Y if we *do not* condition on Y;
- when X and Z collide at Y, the ball passes through Y if we *condition* on Y.

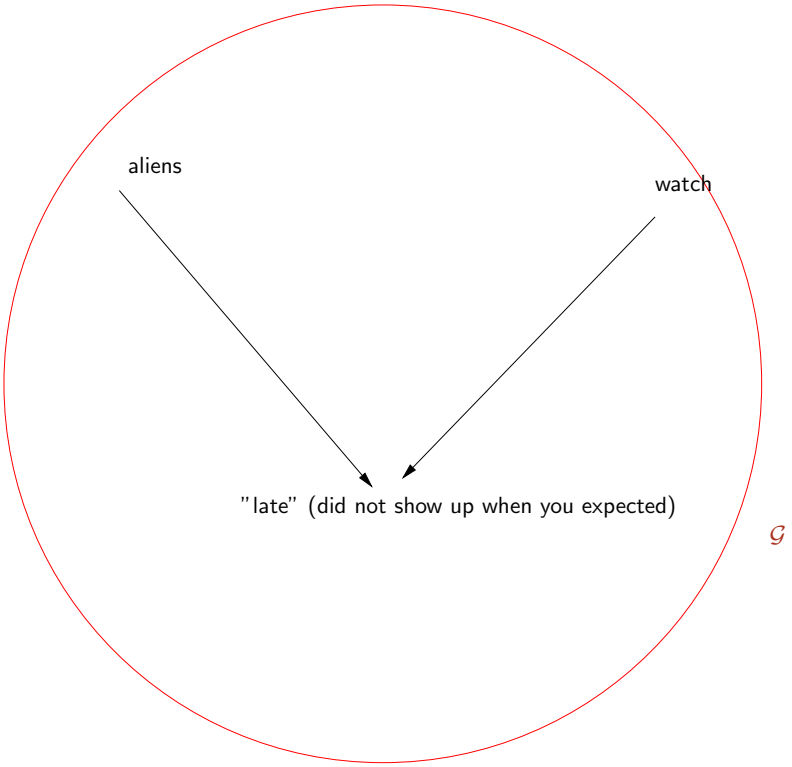
Finally, conditioning on the descendant of a collider has the same effect as conditioning on the collider. For example,



Suppose that X corresponds to *burglary,* Z to *earthquake,* Y to an event where an *alarm* is activated in your building, W to *friend's report* (e.g. friend hears the alarm and calls to tell you).

In general, the chances of a burglary or an earthquake are independent. But, if an alarm goes off in your building, then your suspicions of the cause (either burglary or earthquake) are highly dependent upon conditioning on Y. Suppose now that you do not hear the alarm, but a friend tells you that the alarm went off. In this case, we condition on W and, again, the events of burglary or an earthquake are no longer independent (upon conditioning on Y).

Here is an amusing example:



Your friend appears to be late for a meeting with you. There are two explanations:

- she was abducted by aliens or
- you forgot to set your watch ahead one hour for daylight savings time.

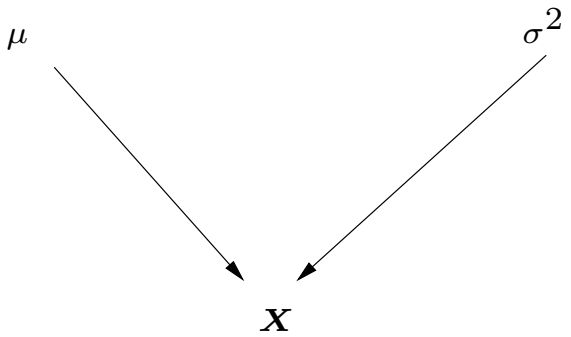
The variables "aliens" and "watch" are blocked by a collider, which implies that they are independent. This is a reasonable model: before we know anything about your friend not showing up when you expected, we would expect these variables to be independent. But, upon learning that she did not show up, "aliens" and "watch" become highly dependent.

Example: Suppose that our measurement vector $\boldsymbol{X} = \boldsymbol{x}$ given μ, σ^2 follows

$$f_{\boldsymbol{X} \,|\, \mu, \Sigma^2}(\boldsymbol{x} \,|\, \mu, \sigma^2)$$

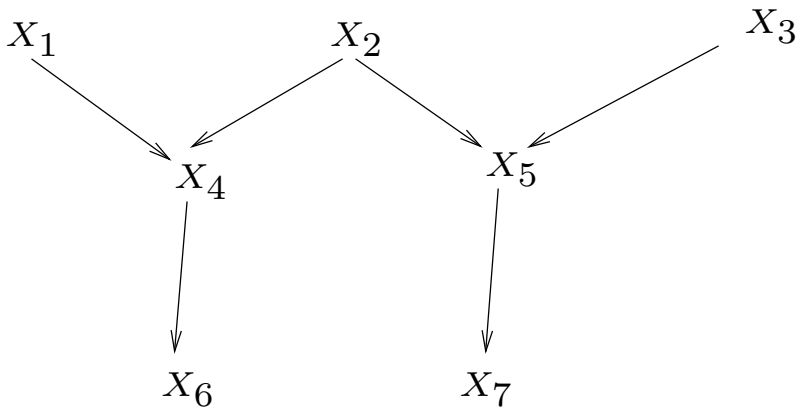
and that we choose independent priors for μ and σ^2:

$$f_{\mu, \Sigma^2}(\mu, \sigma^2) = f_{\mu}(\mu)\, f_{\Sigma^2}(\sigma^2).$$



Here, μ and σ^2 are d-connected given the observations $\boldsymbol{X}$ and, therefore, are *not* conditionally independent given $\boldsymbol{X}$ in general.

Example:



1 and 3 are d-separated (given the empty set $\emptyset$)

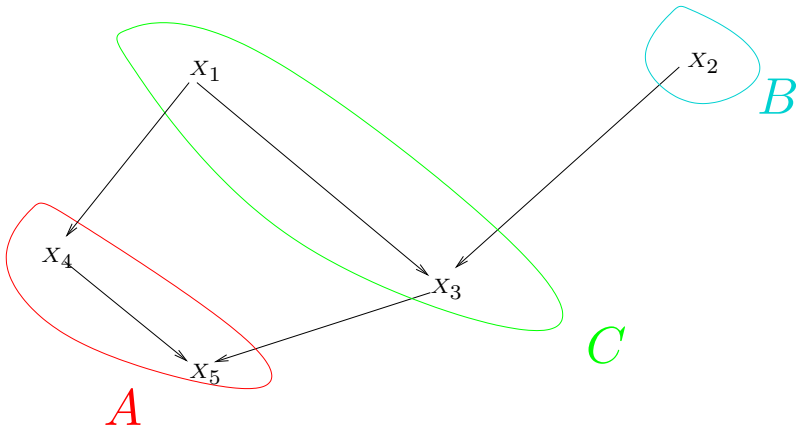
1 and 3 are d-connected given $\{6, 7\}$.

1 and 3 are d-separated given $\{6, 7, 2\}$.

(Back to) an Earlier Example: Recall that we wish to prove

$$X_2 \perp\!\!\!\perp \{X_4, X_5\} \mid \{X_1, X_3\}$$

which we stated on p. 17.



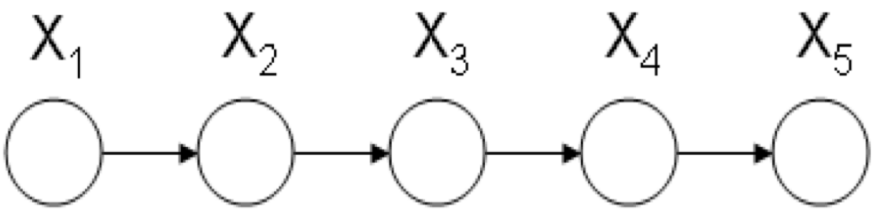
$$A = \{4, 5\}, \quad B = \{2\}, \quad C = \{1, 3\}.$$

Note that

- 2 and 4 are d-separated given C and
- 2 and 5 are d-separated given C

implying that A and B are d-separated given C. Then, Theorem 3 implies that $\{X_4, X_5\} \perp\!\!\!\perp X_2 \,|\, \{X_1, X_3\}$, which completes the proof.

Example: Simple Markov chain graph.



Are the following conditional independence relationships true?

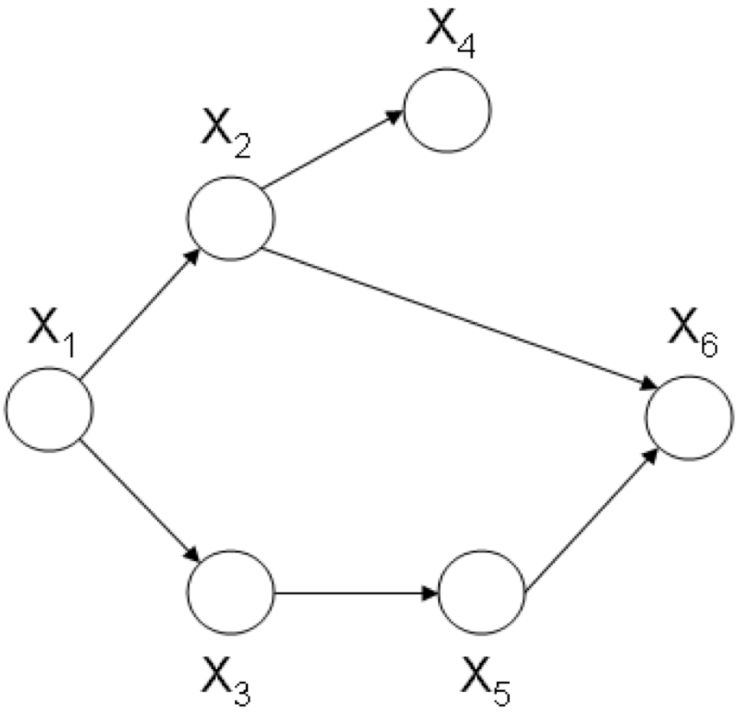
$$X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \tag{6}$$

$$X_1 \perp\!\!\!\perp X_5 \,|\, \{X_3, X_4\}. \tag{7}$$

To determine if (6) is true, we shade node X_2. This blocks balls traveling from X_1 to X_3 and proves that (6) is valid.

Similarly, after shading nodes X_3 and X_4, we find that no ball can travel between X_1 and X_5 and hence (7) holds.

Example:



Are the following conditional independence relationships true?

$$X_4 \perp\!\!\!\perp \{X_1, X_3\} \,|\, X_2 \tag{8}$$

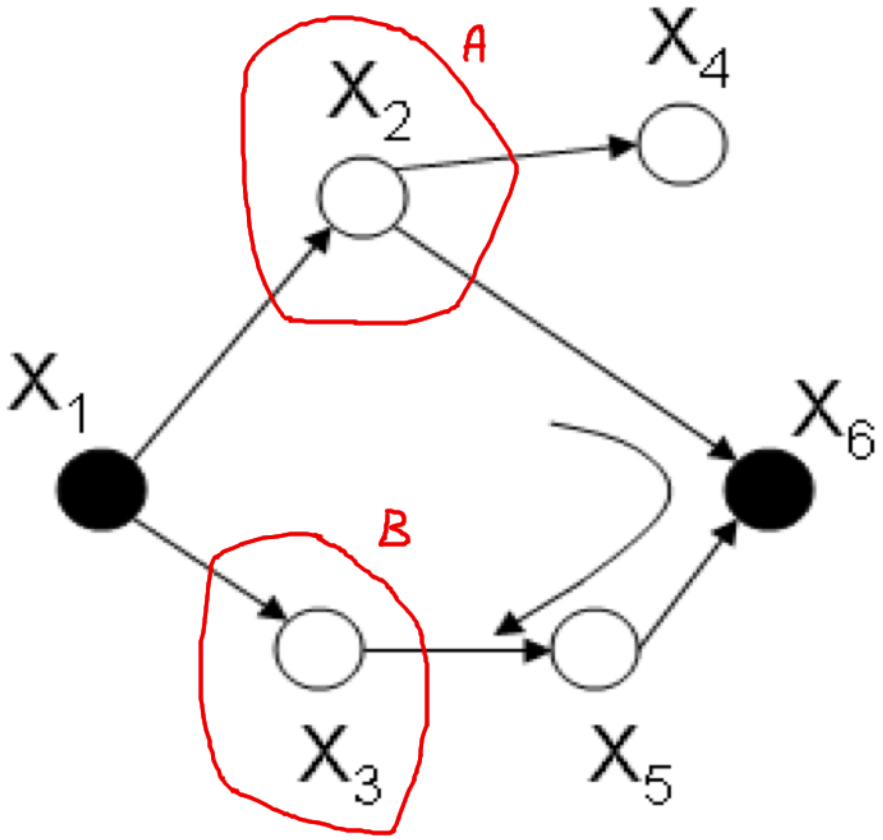
$$X_1 \perp\!\!\!\perp X_6 \,|\, \{X_2, X_3\} \tag{9}$$

$$X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1, X_6\} \tag{10}$$

To prove (8), we must show that $X_4 \perp\!\!\!\perp X_1 \,|\, X_2$ and $X_4 \perp\!\!\!\perp X_3 \,|\, X_2$. Can we find a path for the Bayes ball from X_4 to X_1 once X_2 is shaded? Can we find a path for the Bayes ball from X_4 to X_3 once X_2 is shaded? No, so (8) is true!

Can we find a path for the Bayes ball from X_1 to X_6 once X_2 and X_3 are shaded? No, so (9) is true!

Can we find a path for the Bayes ball from X_2 to X_3 once X_1 and X_6 are shaded?



Yes, so (10) is false!

Markov Equivalent Graphs

Graphs that look different may actually correspond to the same independence relations.

Definition 14. *(A few definitions) Consider a DAG $\mathcal{G}$. We denote by $\mathcal{I}(\mathcal{G})$ all the independence statements implied by $\mathcal{G}$.*

Now, two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ defined over the same random variables V are Markov equivalent if

$$\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2).$$

Given a DAG $\mathcal{G}$, let skeleton($\mathcal{G}$) denote the undirected graph obtained by replacing the arrows with undirected edges.

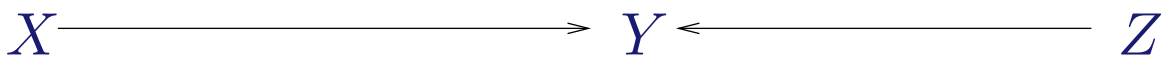
Theorem 4. *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if*

(i) *skeleton($\mathcal{G}_2$) = skeleton($\mathcal{G}_2$) and*

(ii) *$\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders.*

Example: The following three DAGs are Markov equivalent:

But this DAG:



is *not* Markov equivalent to the above three graphs, because condition (ii) in Theorem 4 is not satisfied.

Probability and Undirected Graphs

Definition 15. *An undirected graph $\mathcal{G} = (V, E)$ has a finite set of vertices (nodes) V and a set of edges E that consists of pairs of vertices.*

Definition 16. *A subset $U \subset V$ with all edges connecting the vertices in U is called a subgraph of $\mathcal{G}$.*

Definition 17. *Two vertices X and Y are adjacent if there is an edge between them, and this is written*

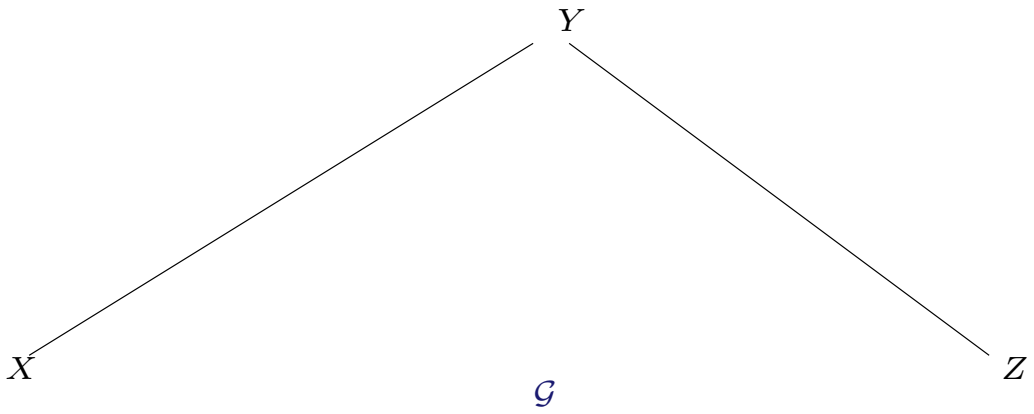
$$X \sim Y.$$

Definition 18. *A graph is called complete if there is an edge between every pair of vertices.*

Definition 19. *A sequence of vertices $X_0, X_1, \ldots, X_n$ is called a path if*

$$X_{i-1} \sim X_i \quad \text{for each } i.$$

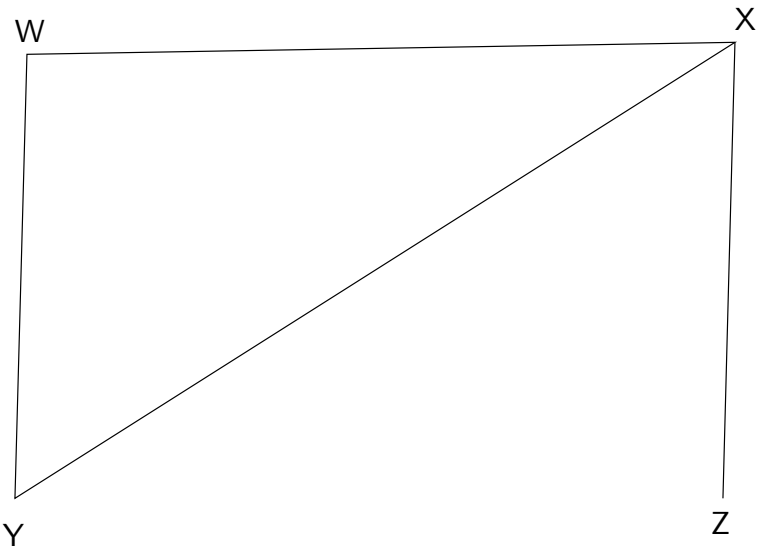
Example:



$V = \{X, Y, Z\}$ and $E = \{(X, Y), (Y, Z)\}$. In undirected graphs, there is no notion of order when defining the edges.

Definition 20. *If A, B, and C are disjoint subsets of V, we say that C separates A and B provided that every path from an X in A to a Y in B contains a vertex in C.*

Example:



$\{Y, W\}$ and $\{Z\}$ are separated by $\{X\}$.

$\{W\}$ and $\{Z\}$ are separated by $\{X\}$.

Definition 21. *(Pairwise Markov) For F a joint distribution of $(X_1, X_2, \ldots, X_K)$, we associate a pairwise-Markov graph $\mathcal{G}$ with F if the following holds:*

do not connect X_i and X_j with an edge if and only if

$$X_i \perp\!\!\!\perp X_j \,|\, X_{\text{rest}}$$

where "rest" refers to all other nodes besides i and j.

Theorem 5. *Let $\mathcal{G}$ be a pairwise Markov graph for F. Let A, B, and C be non-overlapping subsets of V such that C*

separates A and B. Then

$$X_A \perp\!\!\!\perp X_B \,|\, X_C$$

where $X_\cdot = \{X_i \,|\, i \in \cdot\}$.

Here is a short statement of the above theorem:

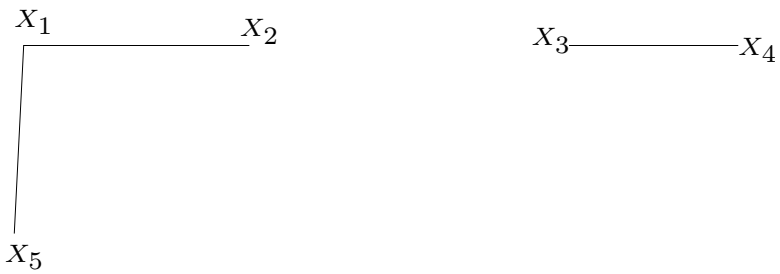
$$X_A \perp\!\!\!\perp X_B \,|\, X_C \quad \Longleftrightarrow \quad C \text{ separates } A \text{ and } B.$$

Remarks:

- If A and B are not connected, we may regard them as "being separated by the empty set." Hence, Theorem 5 implies that $X_A \perp\!\!\!\perp X_B$.

- Theorem 5 defines the "Bayes-ball approach" for undirected graphs. Here, it is straightforward to establish conditional independence.

Example: Suppose that we have a distribution F for $(X_1, X_2, X_3, X_4, X_5)$ with associated pairwise Markov graph:



Then, Theorem 5 implies that

$$(X_1, X_2, X_5) \perp\!\!\!\perp (X_3, X_4) \quad \text{(conditional on nothing)}$$
$$X_2 \perp\!\!\!\perp X_5 \mid X_1.$$

Definition 22. *(Global Markov) For F a joint distribution of $(X_1, X_2, \ldots, X_K)$ and $\mathcal{G}$ an undirected graph, we say that F is globally $\mathcal{G}$ Markov if and only if, for non-overlapping sets A, B, and C,*

$$X_A \perp\!\!\!\perp X_B \,|\, X_C \iff C \text{ separates } A \text{ and } B.$$

The pairwise and global Markov properties are equivalent, i.e.

Theorem 6. *F is globally $\mathcal{G}$ Markov $\iff$ $\mathcal{G}$ is a pairwise Markov graph associated with F.*

Example:

$$X \perp\!\!\!\perp Z \mid Y$$
$$X \perp\!\!\!\perp W \mid Z$$
$$X \perp\!\!\!\perp W \mid Y$$
$$X \perp\!\!\!\perp W \mid (Z, Y).$$

Question

What can we say about the pdf/pmf of $\boldsymbol{X} = [X_1, X_2, \ldots, X_K]^T$ based on an undirected pairwise Markov graph?

Definition 23. *A clique is a set of vertices on a graph that are all adjacent to each other.*

Definition 24. *A clique is maximal if it is not possible to add another vertex to it and still have a clique.*

Definition 25. *Any positive function might be called a potential.*

Result: *Under certain conditions (positivity), a pdf/pmf p for $\boldsymbol{X} = [X_1, X_2, \ldots, X_K]^T$ is globally $\mathcal{G}$ Markov if and only if there exist potentials $\psi_C(x_C)$ such that*

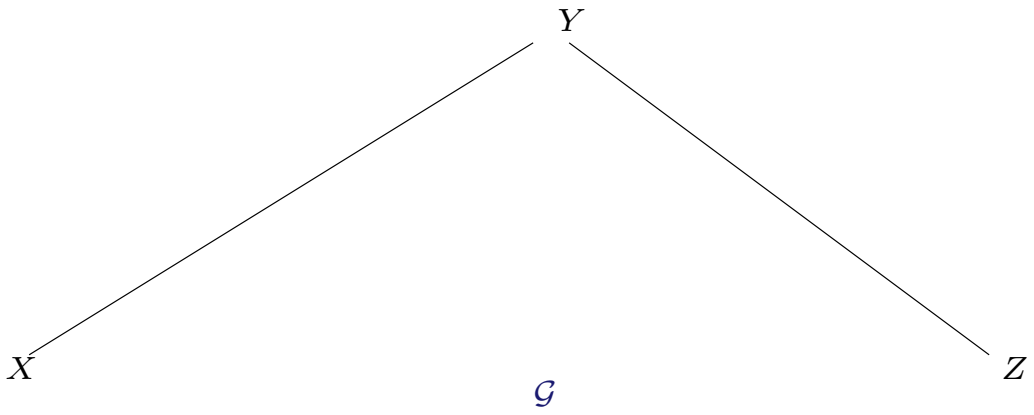
$$p(\boldsymbol{x}) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where $\mathcal{C}$ is the set of maximal cliques. Of course, it does not cost us anything to add more cliques (in addition to the maximal ones), so we can write

$$p(\boldsymbol{x}) \propto \prod_{C \in \mathcal{C}'} \psi_C(x_C)$$

where $\mathcal{C}'$ is the set of all cliques, say.

(Back to) Example:



The maximal cliques in this example are $C_1 = \{X, Y\}$ and $C_2 = \{Y, Z\}$. Hence, under certain conditions, F is globally $\mathcal{G}$ Markov if and only if

$$p(x, y, z) \propto \psi_1(x, y) \cdot \psi_2(y, z)$$

for some positive functions ψ_1 and ψ_2.

Suppose that we know that $p(x, y, z)$ factorizes as

$$p(x, y, z) \propto \psi_1(x, y) \cdot \psi_2(y, z).$$

We can then draw the above graph $\mathcal{G}$ to represent this factorization and conclude by separation properties (say) that

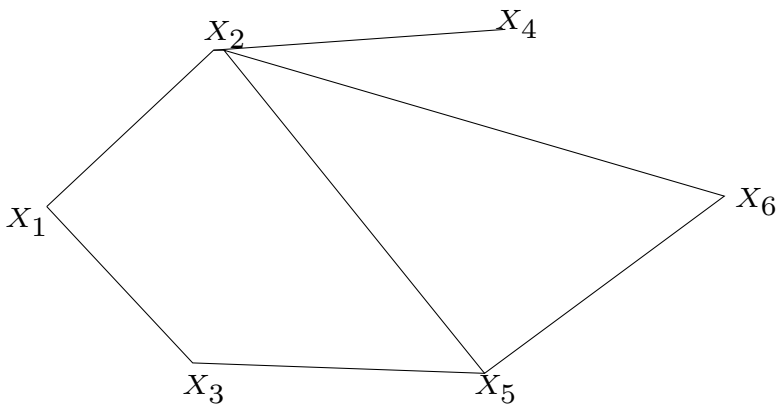
$X \perp\!\!\!\perp Z \,|\, Y$. We can do this analytically as well:

$$p(x \,|\, y, z) \propto p(x, y, z) \propto \psi_1(x, y) \cdot \psi_2(y, z) \propto \psi_1(x, y)$$

implying that

$$p(x \,|\, y, z) = p(x \,|\, y) \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Z \,|\, Y.$$

Example:



Here are the maximal cliques: $\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_5, X_6\}, \{$
$\{X_3, X_5\}$. Hence, under certain conditions, F is globally $\mathcal{G}$ Markov if and only if

$$\begin{aligned} p(\boldsymbol{x}) \quad \propto \quad & \psi_{12}(x_1, x_2) \cdot \psi_{13}(x_1, x_3) \cdot \psi_{24}(x_2, x_4) \cdot \psi_{35}(x_3, x_5) \\ & \cdot \psi_{256}(x_2, x_5, x_6). \end{aligned}$$

Factorization of the Multivariate Gaussian Pdf

Consider a multivariate Gaussian random vector $\boldsymbol{x}$ distributed as $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} \cdot |K|^{1/2} \cdot \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T K (\boldsymbol{x} - \boldsymbol{\mu})\right]$$

where $K = \Sigma^{-1}$ is the *precision matrix* of the distribution.

This Gaussian density factorizes with respect to $\mathcal{G}$ if and only if

$$i \nsim j \quad \Longrightarrow \quad K_{i,j} = 0$$

for $i, j = 1, 2, \ldots, n$. In words: the precision matrix has zero entries for non-adjacent vertices.

Summary

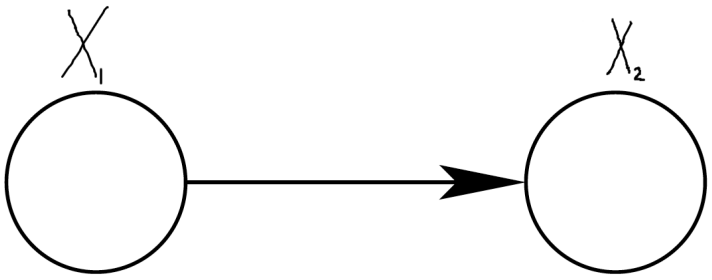
For a Markov graph $\mathcal{G}$ (directed or undirected), the following result holds: if node sets A and B are separated given C, then

$$X_A \perp\!\!\!\perp X_B \,|\, X_C.$$

But, what can we say about conditional dependence of X_A and X_B if A and B are connected given C? **Nothing.** In general

$$A \text{ and } B \text{ are connected given } C \quad \not\Rightarrow \quad X_A \not\perp\!\!\!\perp X_B \,|\, X_C.$$

For example, if this DAG



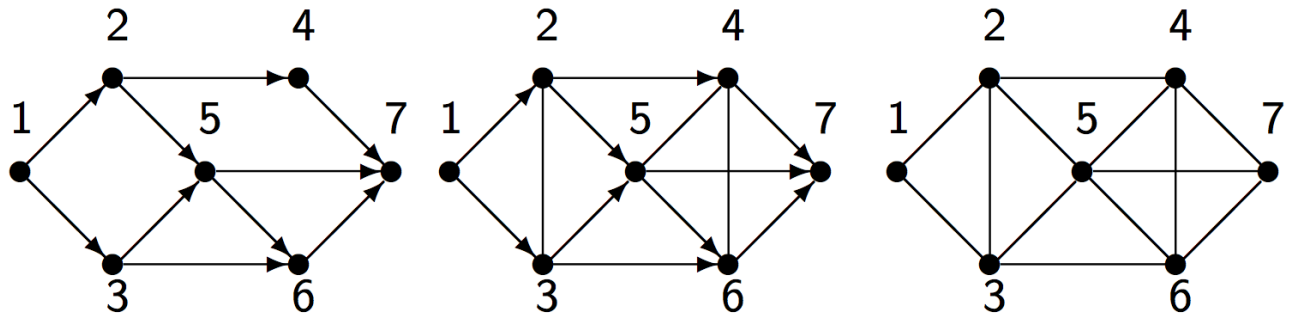
represents a probability distribution, then

$$p(x_1, x_2) = p(x_1)\, p(x_2 \,|\, x_1)$$

but we have complete freedom to choose $p(x_2 \,|\, x_1)$. If we choose $p(x_2 \,|\, x_1) = p(x_2)$, then X_1 and X_2 are independent!

Moralization: Conversion of DAGs to Undirected Graphs

The *moral graph* $\mathcal{G}^{\mathrm{m}}$ of a DAG $\mathcal{G}$ is obtained by adding undirected edges between unmarried parents (i.e. joining or "marrying" parents of unshielded colliders) and subsequently dropping directions, as in the example below:



Proposition. *If F factorizes with respect to $\mathcal{G}$ (i.e. $\mathcal{G}$ represents F), then it factorizes with respect to its moral graph $\mathcal{G}^{\mathrm{m}}$.*

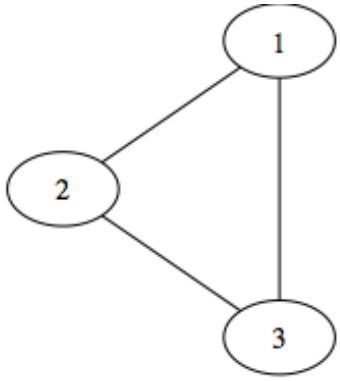
This is seen directly from the factorization:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i \in V} p_{X_i \,|\, X_{\mathrm{pa}_i}}(x_i \,|\, x_{\mathrm{pa}_i}) \propto \prod_{i \in V} \psi_{\{i\} \cup \mathrm{pa}_i}(x_{\{i\} \cup \mathrm{pa}_i})$$

since $\{i\} \cup \mathrm{pa}_i$ are all cliques in $\mathcal{G}^{\mathrm{m}}$.

A Bit About Factor Graphs (which are Popular in Coding Theory)

Motivation: So far, our focus has been on conditional-independence statements that are represented by a graph $\mathcal{G}$. What if we wish to represent pdf/pmf factorization? Consider the following graph:



At the first glance, we see a 3-clique and we can only give the following (totally noninformative) representation of the corresponding distribution:

$$p(\boldsymbol{x}) \propto \psi_{123}(x_1, x_2, x_3) \qquad \text{Model (a)} \qquad (11)$$

but, suppose that we know that there exist only pairwise interactions; then a special case of (11) which takes this knowledge into account is:

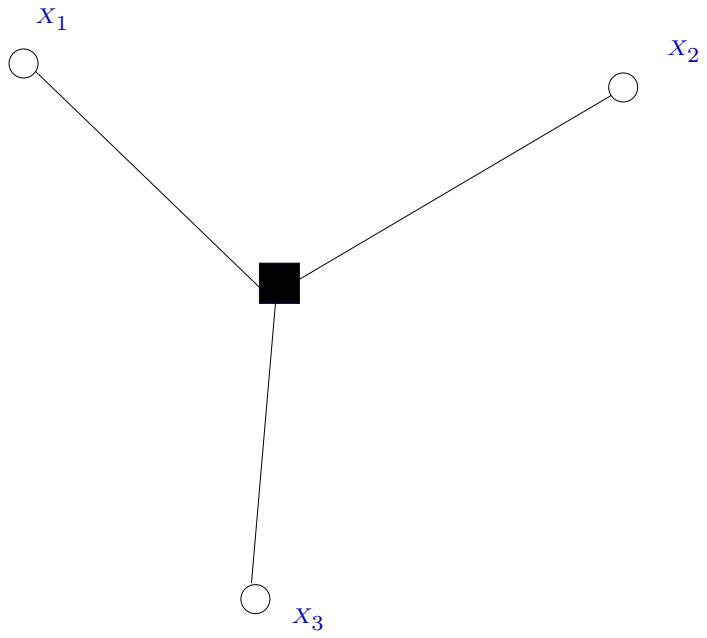
$$p(\boldsymbol{x}) \propto \psi_{12}(x_1, x_2) \cdot \psi_{23}(x_2, x_3) \cdot \psi_{13}(x_1, x_3) \qquad \text{Model (b)}$$

but its undirected-graph representation is the same 3-clique!

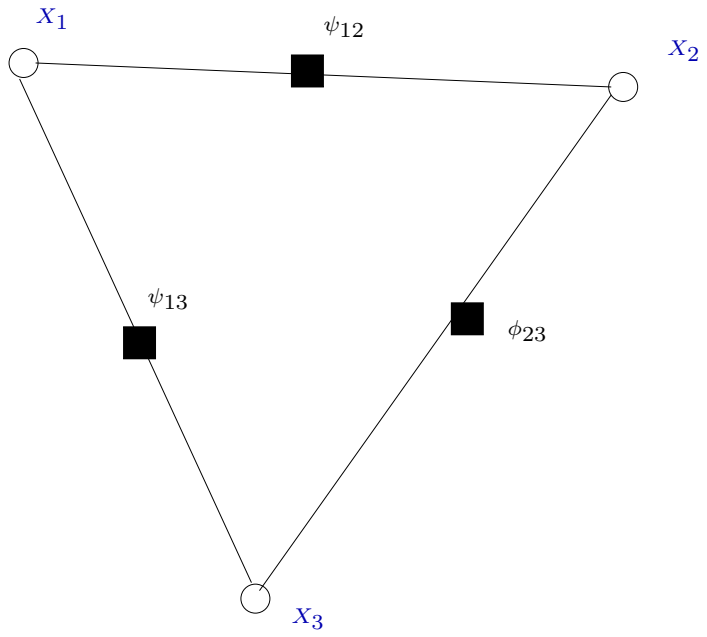
Add a new *factor node* for every product term (factor) in the pdf/pmf representations. Connect the factor boxes with the variables that they "touch." (Hence, in a factor graph, edges exist only between the variable nodes and factor nodes.)

Here are factor graphs:

for model (a)



and for model (b)



A bit more rigorously, we can say that the ingredients of factor graphs are

- $V = \{1, 2, \ldots, N\} \equiv$ set of vertices depicting random variables;
- $\Psi = \{a, b, c, \ldots\} \equiv$ index sets of factors;
- $E \equiv$ set of edges

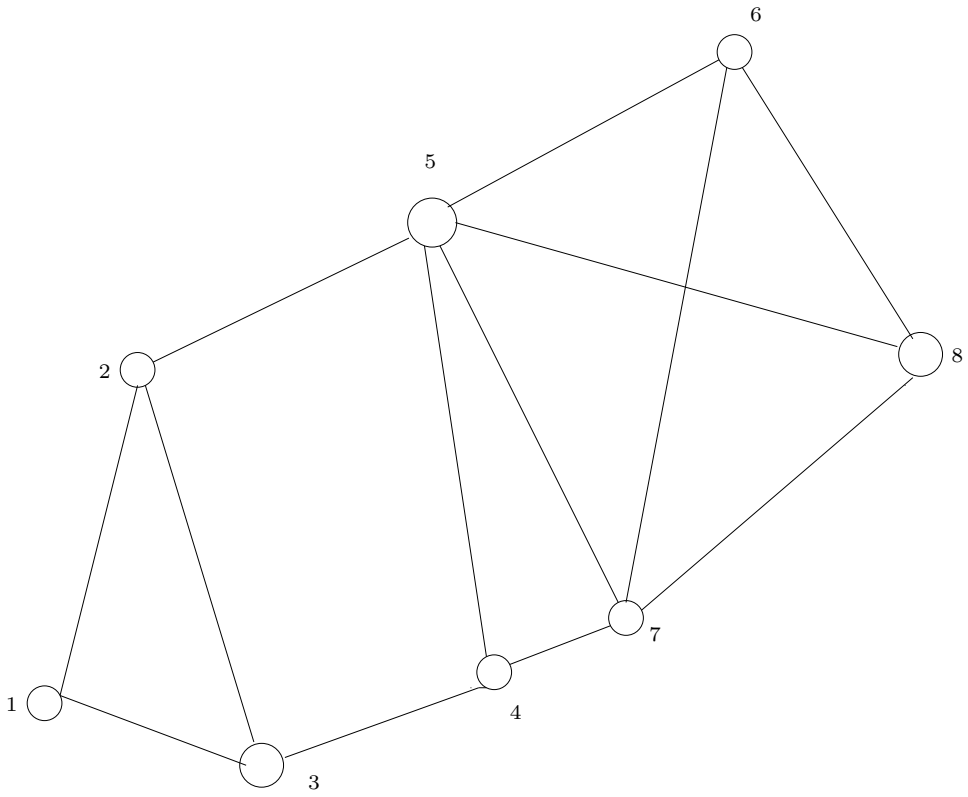
describing the factorization

$$p(\boldsymbol{x}) \propto \prod_{a \in \Psi} \psi_a(\boldsymbol{x}_a).$$

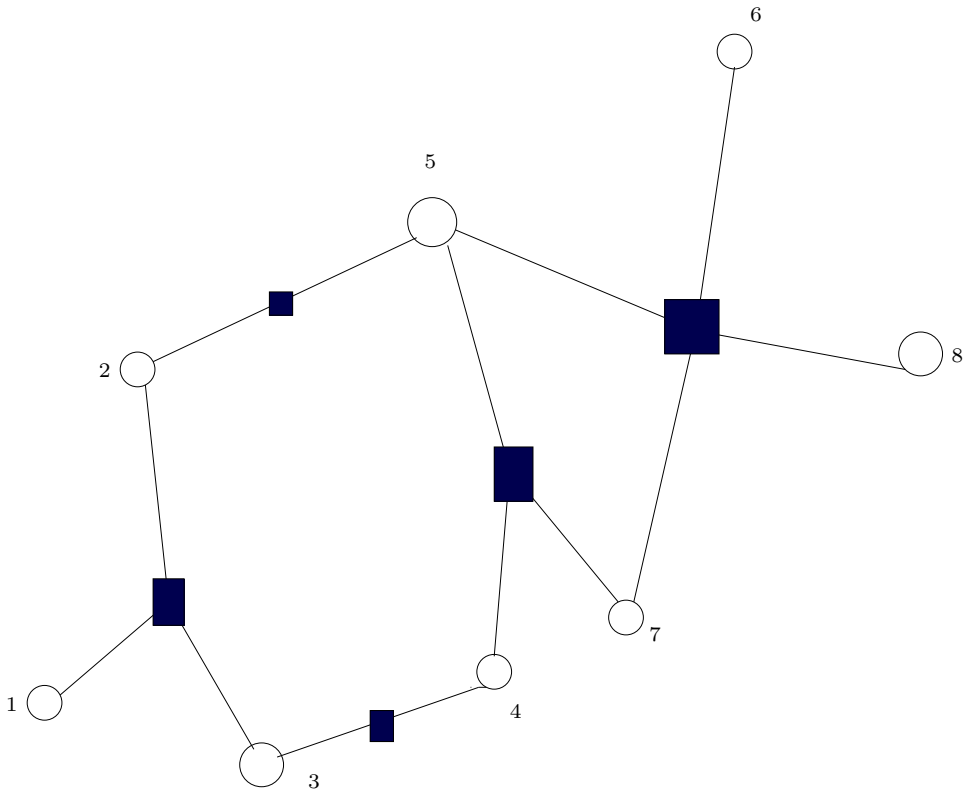
(Recall that $\boldsymbol{X}_a = \{X_i \,|\, i \in a\}$.)

Any directed or undirected graph can be converted into a factor graph.

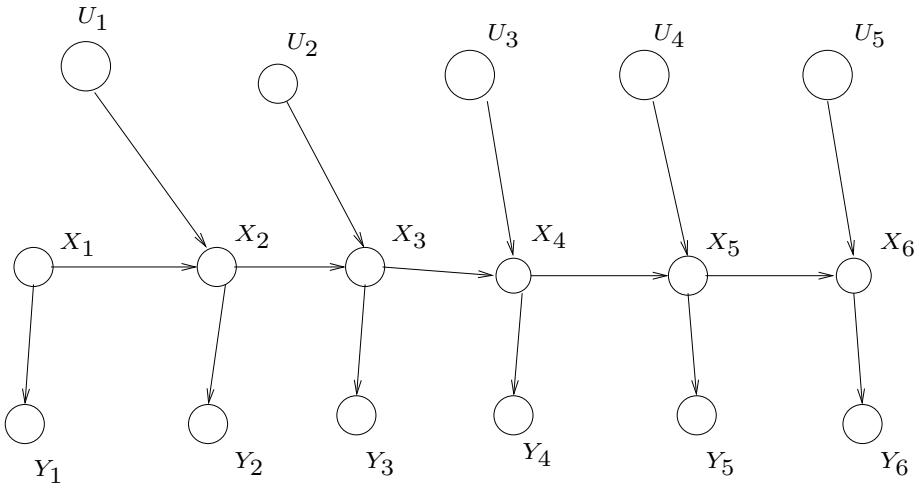
Example: An undirected graphical model:



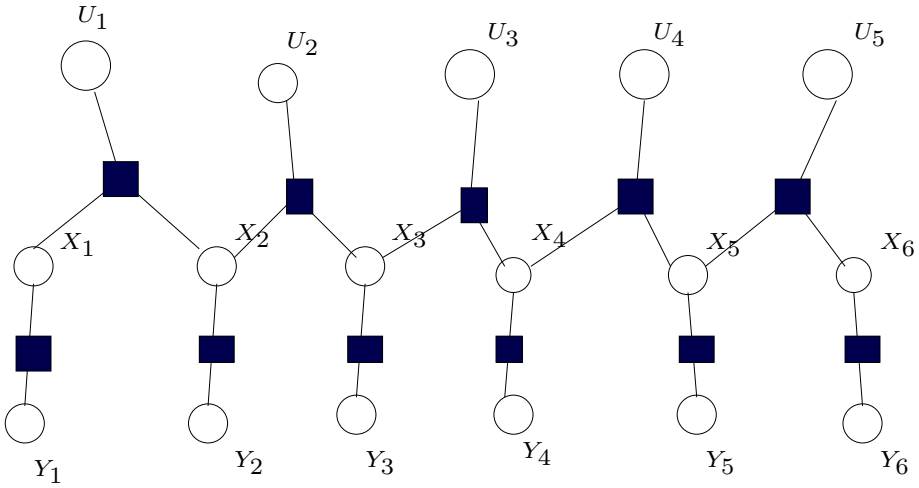
and its (possible) factor graph:



Example: A directed graphical model:



and its factor graph:



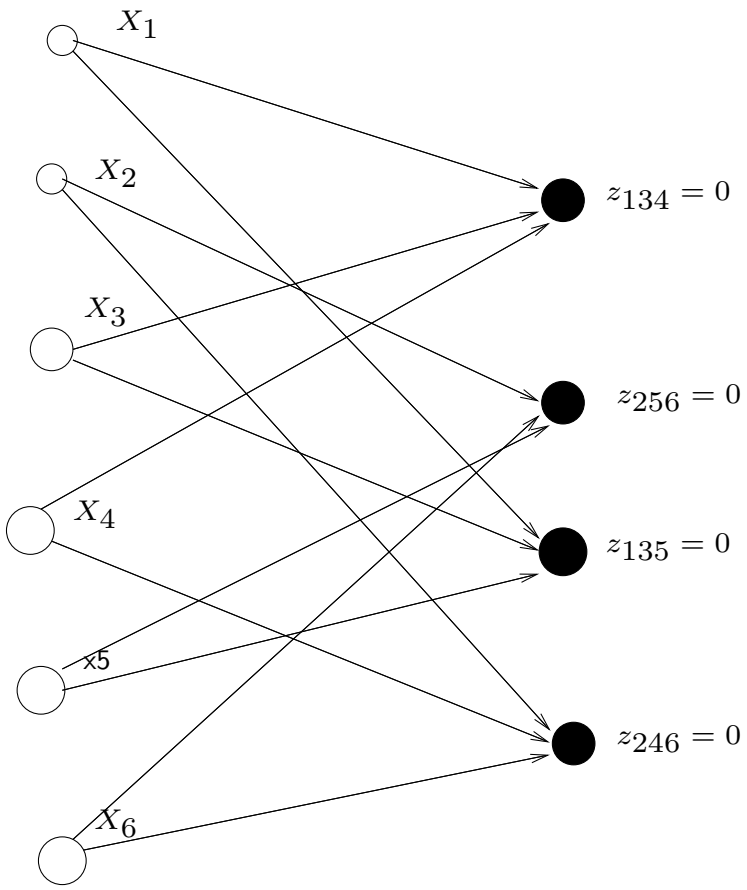
Belief-propagation algorithms can be derived for factor graphs. This topic will not be discussed here, but understanding the basic belief-propagation algorithm for undirected tree graphs is key to understanding its version for *factor trees* (i.e. *factor graphs that have no loops*). Unlike the basic belief-propagation algorithm (covered later in this class), its version for factor trees has two types of messages: messages from variable to factor nodes and messages from factor nodes to variable nodes.

Example: Application of Graphical Models to Coding Theory

An example, roughly taken from

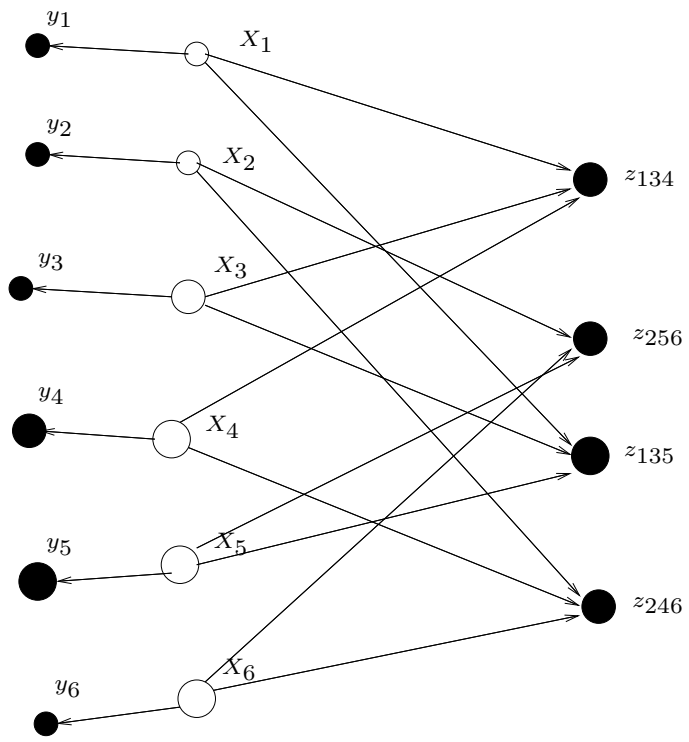
(Wainwright & Jordan 03) M.J. Wainwright and M.I. Jordan, "Graphical models, exponential families, and variational inference," Report no. 649, Department of Statistics, University of California, Berkeley, CA, 2003.

Consider this DAG representation of a small parity-check code:



where $X_i \in \{0, 1\}$, $i = 1, 2, \ldots, 6$.

The code is defined by setting each *parity variable* $z_{s,t,u}$, $(s,t,u) \in \{\{1,3,4\},\{1,3,5\},\{2,5,6\},\{2,4,6\}\}$ to zero. Hence, the variables $z_{s,t,u}(s,t,u) \in \{\{1,3,4\},\{1,3,5\},\{2,5,6\},\{2,4,6\}\}$ are "observed," which is why they are shaded. Also, the pmf $p(z_{134} \,|\, x_1, x_3, x_4)$ (say) is simply the pmf table describing the $x_1 \oplus x_3 \oplus x_4$ operation. Now, suppose that the random variables $X_1, X_2, \ldots, X_6$ are *hidden* and that we observe only their noisy realizations $y_1, y_2, \ldots, y_6$:



Then, our decoding problem can be posed as determining the marginal posterior pdfs

$$p(x_i \,|\, y_1, y_2, y_3, y_4, y_5, y_6, z_{134} = 0, z_{256} = 0, z_{135} = 0, z_{246} = 0)$$

for $i = 1, 2, \ldots, 6$.