MAC6916 PROBABILISTIC GRAPHICAL MODELS LECTURE 4: MARKOV EQUIVALENCE AND MARKOV NETWORKS

DENIS D. MAUÁ

1. Motivation

In this lecture, we discuss the classes of graphs that are equivalent under d-separation, and we examine the representation of independency by undirected graphs, resulting in the concept of Markov networks. We also compare Bayesian networks and Markov networks as representational devices for independencies.

2. Equivalence

Under the semantics of d-separation, an acyclic directed graph can be seen as a compact representation of an independency relation. As we discussed previously, not every independency relation can be represented in such a way (because d-separation satisfy properties that not every distribution satisfies). Pearl advocated independecy (and their corresponding graphical representation) as a more fundamental knowledge than probability. In this view, independencies may exist without any reference to a probability function. Indeed, the d-separation semantics and its variants can be used to represent independencies in many non-probabilistic theories such as Dempster-Shafer theory [1] and credal networks (interval-valued probability models) [2]. Notably, an independency relation admits many graphical representations. This is clear from the definition of d-separation in serial and divergent connections. The following discussion aims at characterizing the classes of equivalent structures, and studying their consequences on the representation of independency.

Definition 1. The acyclic directed graphs G and H are *equivalent* if their node set is the same, and equivalence

$$\mathcal{X} \perp_d^G \mathcal{Y} \mid \mathcal{Z} \text{ if and only if } \mathcal{X} \perp_d^H \mathcal{Y} \mid \mathcal{Z}$$

for every subset of nodes $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$.

Equivalence is also called I-equivalence (from independency-equivalence), and Markov equivalence. The latter is due to the fact that the local and global Markov properties are equivalence, hence, two graphs are equivalent if and only if the set of Markov properties of either graph is satisfied by the other one. The three directed graphs in Figure 1 are equivalent.

Consider a graph G with an arc $X \to Y$. Any equivalent graph H must have either the same arc or a reversed arc $Y \to X$, since two variables connected by an arc

Date: March 16, 2016.

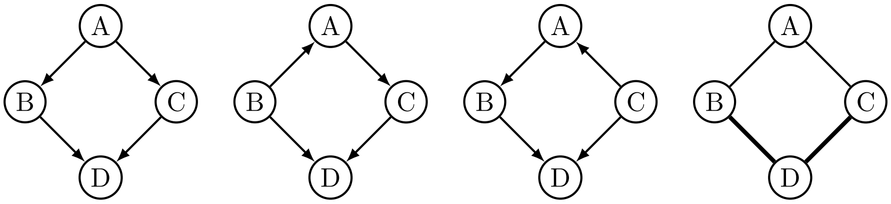


Figure 1. Three equivalent graphs and their skeleton with the common v-structure highlighted.

are d-connected, and any two variables not connected by an arc can be d-separated by some set of variables (including the empty set). This implies that equivalent graph must have the same underlying undirected structure. This is formalized by the concept of skeleton:

skeleton

Definition 2. The *skeleton* of a directed graph $G = (V, E)$ is the undirected graph $H = (V, F)$ such that $X \to Y \in E$ if and only if $X - Y \in F$.

It follows from our previous discussion that candidate equivalent graphs must have the same skeleton. For example, the rightmost graph in Fig 1 is the skeleton of the directed graphs to the left. Now consider 3-node graphs G and H, such that G is a serial or divergent connection and H is a convergent connection. While the two graphs have the same skeleton, they represent different d-separation relations (and are thus not equivalent). This is formalized by the concept of v-structure:

immorality

Definition 3. An *immorality* is a triple of nodes X, Y, Z such that $X \to Y$, $Z \to Y$ and X and Z are non-adjacent (i.e., there is no arc connecting X and Z).

immorality

An *immorality* is also called a v-structure (although this name is also used to describe a general convergent connection). By a previous reasoning, two equivalent graphs must have the same set of immoralities lest we can find a triple (one that is a serial or divergent connection in one and a convergent connection in the other) that induces different d-separations. Thus we have that two graphs are equivalent only if they have the same skeleton and the same set of immoralities. Verma and Pearl [3] showed that the converse is also true .

Theorem 1. *Two acyclic directed graphs are equivalent if and only if they have the same skeleton and the same set of immoralities.*

The graphs in Figure 1 all satisfy the theorem above. The arcs of the single common immorality $B \to D \leftarrow C$ appear highlighted in their skeleton. The set of acyclic directed graphs can be partitioned into its sets of equivalent graph. According to the previous theorem, each part can be uniquely represented by the skeleton and the list of immoralities. Notice that, as for d-separation, the equivalence of graphs is a graphical property and can be tested efficiently (in linear time in the number of arcs).

essential graph

Definition 4. The *essential graph* of a class $\mathcal{G}$ of equivalent acyclic directed graphs $G = (V, E_G)$ is the graph $H = (V, A)$, where $A = \cup_{G \in \mathcal{G}} E_G$, that is, A contains an arc $X \to Y$ if and only if there is a graph G in $\mathcal{G}$ with that arc.

Since all the graphs in an equivalence class have the same skeleton and set of immoralities, the essential graph is unique. A closely related concept is obtained by

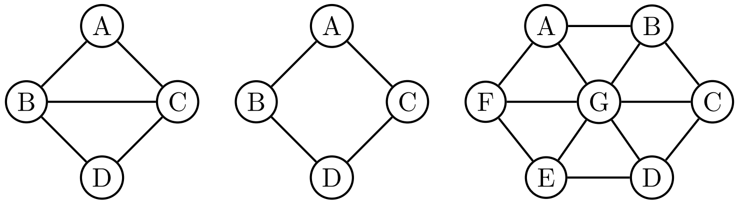


Figure 2. A non-chordal graph (left) and a non chordal graphs (center and right).

interpreting the essential graph as graph that contain both undirected and directed edges. Call any symmetric relation $X \to Y$ iff $Y \to X$ an *undirected edge*, and any asymmetric relation $X \to Y$ iff $X \not\to Y$ a *directed edge*. A *partially directed graph* is a graph that contains directed and undirected edges.

undirected edge
directed edge
partially directed graph

Definition 5. The partially directed graph of an equivalence class $\mathcal{G}$ is the (partially) directed graph that has the same skeleton as a graph in $\mathcal{G}$ and has a directed arc $X \to Y$ if and only this arc appears in all graphs in the class.

The partially directed graph represents possible orientations of the arcs in an equivalence class. Any orientation of the undirected graphs that is topologically consistent and does not introduce new immoralities leads to a graph that represents the same set of d-separations. A partially directed graph is a *chain graph* if there is no cycle that uses only directed edges. The partially directed graph of an equivalence is a chain graph. It is possible to modify d-separation to work directly on chain graphs.

chain graph

Given the equivalence between the Markov properties encoded by a graph and the factorization of a distribution, the graphs in an equivalence class can be seen as different *parametrizations* of a probability distribution.

parametrization

3. A Few Concepts From Graph Theory

We defined an undirected graph as a directed graph $G = (V, E)$ where E is symmetric. When working exclusively with undirected graphs is more convenient to define undirected graphs as a pair (V, E) where E is a set of *unordered pairs* of vertices. We will adopt this definition in the following and denote an (undirected) edge as $X - Y$. A *loop* (or undirected cycle) is a trail that starts and ends with the same node. A graph with no loops is a tree (even if it is not connected).

loop

Definition 6. A graph is *chordal* if every loop has a subsequence which is a triangle (i.e., 3-node loop).

chordal graph

Figure 2 shows a chordal graph in the left (any permutation of the loop A, B, C, D has loops A, B, C or B, C, D) and non-chordal graphs in the center and right (e.g., in the graph in the center, the loop A, B, C, D has subsequence which is also a loop). Note that the rightmost graph is non-chordal even though is made up of triangles. The outer loop A, B, C, D, E, F contains no triangles.

Definition 7. A *clique* is a set of nodes which are pairwise connected. A clique is *maximal* if it is not a proper subset of any other clique.

clique
maximal clique

The maximal cliques of the left graph in Fig 2 are $\{A, B, C\}$ and $\{B, C, D\}$ and the non maximal cliques are the edges, the singletons containing each node and the empty set. The maximal clique of the graph in the center is $\{A, B, C, D\}$.

4. Markov networks

We have seen that even though independency is a symmetric concept, it can be well captured by an acyclic directed graph through the semantics of d-separation. Moreover, we saw that d-separation is equivalent to m-separation, thus connecting directed and undirected representations. The m-separation representation, obtained through the relevant moral graph, is *dynamic* in that it depends on the independency (or separation) query. In this section we will look at a undirected representation of independency that is *static*, that is, it is represented by a fixed graph.

u-separation

Definition 8. The sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are *u-separated* by $\mathcal{Z}$, written $\mathcal{X} \perp_u \mathcal{Y} \mid \mathcal{Z}$ if $\mathcal{X}$ and $\mathcal{Y}$ are separated in $G - \mathcal{Z}$.

Hence, m-separation is simply u-separation in the corresponding relevant moral graph. Note that conditioning can only separate sets, but never connect them (this is the static character of u-separation).

Markov network

Definition 9. A *Markov network* is a pair (G, p), where G is an undirected graph over variables $\mathcal{V}$ and $p(\mathcal{V})$ is a joint distribution for $\mathcal{V}$ such that

$$\mathcal{X} \perp_u \mathcal{Y} \mid \mathcal{Z} \text{ only if } \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}\,, \qquad \text{[global Markov property]}$$

for any subsets $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \subseteq \mathcal{V}$.

Consider the graph in the center in Figure 2; it encodes the u-separations

$$A \perp_u D \mid B, C \text{ and } B \perp_u D \mid A, C\,.$$

potential

Definition 10. A *potential* is a nonnegative real-valued function over the joint domain of some set of variables.

factor

A potential is also often called a *factor*. Note that a joint distribution is a potential (and the converse is not generally true). The parametrization of a Markov network in terms of potentials follows from the following result.

Theorem 2 (Hammersley-Clifford Theorem). *Let $\mathcal{C}_1, \ldots, \mathcal{C}_m$ be the maximal cliques of a Markov network structure G. If $p(\mathcal{V}) > 0$ then the global Markov property implies*

$$p(\mathcal{V}) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j(\mathcal{C}_j)\,,$$

where

$$Z = \sum_{\mathcal{V}} \prod_{j=1}^{m} \phi_j(\mathcal{C}_j)$$

partition function

is a normalizing term called the partition function.

clique potentials

The potentials ϕ_c are called *clique potentials*. It is possible to construct a non-positive distribution that satisfies all the global Markov properties in the graph and yet does not factorize over the cliques. The converse result, however, holds even for non-positive distributions.

Theorem 3. *If $p(\mathcal{V})$ is a joint distribution that factorizes as*

$$p(\mathcal{V}) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j(\mathcal{C}_j)\,,$$

where $\mathcal{C}_j$ are the maximal cliques of a graph G, then (G, p) satisfies all the global Markov properties.

The result above uses the following fact about independecy:

Proposition 1. *The sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given the set of variables $\mathcal{Z}$ if and only if there are potentials $\phi_1(\mathcal{X}, \mathcal{Z})$ and $\phi_2(\mathcal{Y}, \mathcal{Z})$ such that*

$$p(\mathcal{X}, \mathcal{Y}) = \phi_1(\mathcal{X}, \mathcal{Z})\phi_2(\mathcal{Y}, \mathcal{Z})\,.$$

It is not difficult to show that Global Markov properties imply

$$X \perp_u \mathcal{V} - \neq X \mid \neq X\,, \qquad \text{[local Markov property]}$$

which in turn imply

X is not adjacent to Y only if $X \perp_u Y \mid \mathcal{V} - \{X, Y\}$. [pairwise Markov property]

Proposition 2. *If $p(\mathcal{V}) > 0$ is a distribution satisfying the pairwise Markov properties, then it also satisfies the global Markov properties.*

Note that unlike Bayesian networks, we defined Markov networks satisfying the global Markov properties, and we deduced the local Markov properties. According to the result above, for positive distributions this choice is arbitrary: we could have defined Markov networks as satisfying the local Markov networks, as we did with Bayesian networks. However, there are non-negative distributions which satisfy the pairwise Markov properties and not the local Markov properties, and distributions which satisfy the local Markov properties and not the global ones. When the distribution is positive, is also common to parametrize it using a distribution from the exponential family:

$$p(\mathcal{V}) = \frac{1}{Z} \exp\left(-\sum_j \psi_j(\mathcal{C}_j)\right) = \exp\left(-\sum_j \psi_j(\mathcal{C}_j) - \ln(Z)\right)\,, \tag{1}$$

where ψ_c are clique potentials.

Definition 11. A *pairwise Markov network* is a Markov network whose maximal cliques are the edges of the graph. It is commonly parametrized as

pairwise Markov network

$$p(\mathcal{V}) = \exp\left(-\sum_{X \in V} \psi(X) - \sum_{X - Y} \psi(X, Y) - \ln(Z)\right)\,,$$

where $\psi(X)$ and $\psi(X, Y)$ are called the node and edge potentials, respectively.

Example 1. Pairwise Markov networks are popular models for image processing tasks. For example, in image segmentation one is interested in classifying each pixel in an image according to pre-determined set of labels (e.g., foreground, background). Every pixel is classified according to local characteristics (pixel intensity, color, etc), and a geometric smoothing is enforced (neighboring pixels should be classified alike). This model can be represented as a pairwise Markov where every pixel is a node and two nodes are adjacent if they correspond to adjacent pixels. The graph

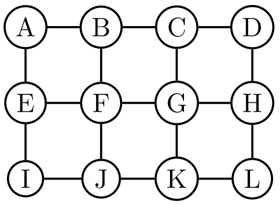


Figure 3. A pairwise Markov network.

in Figure 3 is a pairwise (possibly) representing a 3-by-4 pixels image. The node potentials encode the propensity of each pixel to belong to some specific class, and the edge potentials encode the smoothness (e.g., that neighboring nodes need to be classified equally).

L
restriction

Let ν be some valuation. A *restriction* of a potential $\phi(\mathcal{X})$ with respect to $\mathcal{Y}$ is the potential $\psi(\mathcal{X}-\mathcal{Y}) = \phi \downarrow (\nu, \mathcal{Y})$ such that $\psi\nu' = \phi\nu$ for every valuation ν that agrees with ν on $\mathcal{Y}$.

When convenient we write $\phi(\mathcal{X}-\mathcal{Y}, \mathcal{Y} = \nu(\mathcal{Y})$ to denote the restriction $\phi(\mathcal{X}) \downarrow (\nu, \mathcal{Y})$.

Example 2. Let ν be a valuation with $\nu(Y) = 1$. The restriction of the potential

X	Y	$\phi(X,Y)$
0	0	0.1
0	1	100
1	0	5.2
1	1	0

with respect to ν and Y is

X	$\phi(X, Y = 1)$
0	100
1	0

Factorizing Markov networks are closed with respect to restrictions:

reduced Markov network

Proposition 3. *If (G,p) is a Markov network over $\mathcal{V}$ such that p factorizes as $\prod_j \phi_j(\mathcal{C}_j)/Z$, ν is a valuation and $\mathcal{Y} \subseteq \mathcal{V}$ then $(G-\mathcal{Y}, q)$ is a Markov network over $\mathcal{V}-\mathcal{Y}$ and q factorizes as $\prod_j [\phi_j(\mathcal{C}_j) \downarrow (\nu,\mathcal{Y})]/Z(\nu,\mathcal{Y})$. Moreover, $q(\mathcal{V}-\mathcal{Y}) = p(\mathcal{V}|\mathcal{Y} = \nu(\mathcal{Y}))$. This network is called a* reduced Markov network.

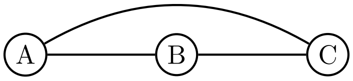
The notation $Z(\nu, \mathcal{Y})$ emphasizes that the partition function of the reduced network depends on the restriction. Hence, computing conditional probabilities in Markov networks is the same as restricting every potential with respect to the evidence and then computing a marginal probability (which requires computing the partition function). Note that the maximal cliques in the reduced network may be smaller than in the original network, so that more than a potential must be combined to form the clique potentials of the reduced network.

Example 3. Consider a Markov network with the graph in the left in Figure 2, and its restriction by $B = 1, C = 0$. The corresponding reduced network is an empty graph over nodes A and D, and the clique potentials are $\phi_1(A) = \phi(A, B = 1)\phi(A, C = 0)$ and $\phi_2(D) = \phi(B = 1, D)\phi(C = 0, D)$. Note that the potential $\phi(B, C)$ is discarded (since it is a constant it is also considered in the partition function). Alternatively, $\phi(B = 1, C = 0)$ could be incorporated into any clique potential (or all of them). The reduced network distribution is $p(A, D|B = 1, C = 0)$.

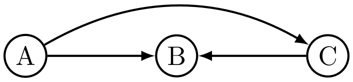
5. Markov nets and Bayesian nets

Since any Bayesian network factorizes (even the ones with zero values), a Bayesian network (G, p) specifies a Markov network $(M[G], p)$ whose partition function is $Z = 1$. Hence, from the parametrization point-of-view, Bayesian networks are subclass of Markov networks with the special property that the partition function is one. However, Bayesian networks can represent dynamic independencies (i.e., independencies created by conditioning), whereas Markov network can represent only static dependencies. So, for example, no Markov network can represent the independencies in a convergent connection. There are independency relations that can be represented by a Markov network but not by any Bayesian network. This is the case, e.g., of the network in the center in Figure 2. Any attempt of representing that relation as a directed graph will either result in creating a convergent connection which misses some independency and introduces others, or lack a direct influence (again, misrepresenting independencies). Thus, from the representational point-of-view, Bayesian networks and Markov networks are incomparable.

What types of independency relations can be represented either by a directed or an undirected graph? The answer follows from our translation from Bayesian networks to Markov networks. Recall that the equivalence class of a Bayesian networks is characterized by the skeleton and the set of immoralities. A Bayesian network without immoralities is thus represented as an undirected graph, in which case d-separation (with any consistent orientation of the arcs) and u-separation coincide. So the class of Bayesian networks such that $G = M[G]$ is contained in the class of Markov networks. What about the converse? Any chordless loop contains independencies that cannot be represented by a directed graph (consider the graph in the center in Figure 2). On the other hand, any triangle



can be represented either as the directed structure



or by some homomorphic structure. It is not difficult to show that the class of Markov networks that can be represented as Bayesian networks is exactly the class of *chordal graphs*; any non-chordal graph cannot be represented (in terms of independencies) by a Bayesian network.

chordal graphs

We will see later that chordal graphs, the intersection of Markov networks and Bayesian networks are an important family of probabilistic graphical models, one in which inference complexity can be graphically characterized. Chordal graphs have the important property that their cliques be arranged in the form of a particular tree, and parametrized as

$$p(\mathcal{V}) = \frac{\prod_{j=1}^{m} p(\mathcal{C}_j)}{\prod_{j=1}^{m-1} p(\mathcal{S}_j)} ,$$

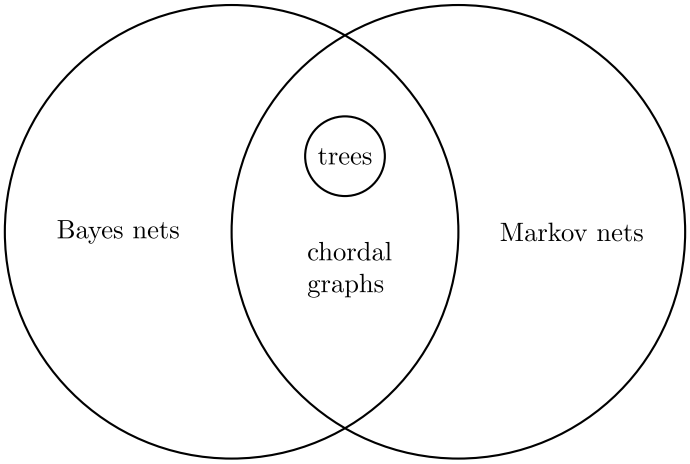


FIGURE 4. Representational power of different formalisms.

where $\mathcal{C}_j$ are the cliques of the graph and $\mathcal{S}_j$ are the intersection of adjacent cliques known as separation sets.

Since the restriction of potentials of a Markov network is also a Markov network, and a Bayesian network parametrizes a Markov network, we can compute conditional probabilities in Bayesian networks by inference in the corresponding restricted Markov network. Let $\mathcal{Y} = y$ be an evidence, and $\mathcal{X}$ be some target variables. To compute $p(\mathcal{X}|\mathcal{Y} = y)$ in a Bayesian network with graph G, obtain the potentials $\phi_i = p(X_i|\text{pa}(X_i)) \downarrow [\mathcal{Y} = y]$ and compute the marginal $q(\mathcal{X})$ is the Markov network defined by the factorization $q(\mathcal{V} - \mathcal{Y}) = Z^{-1} \prod_i \phi_i$. The partition function of this network encodes $Z = p(\mathcal{Y} = y)$, that is, the probability of evidence. To make things computationally more efficient, we can restrict the Markov network to non-barren nodes of $p(\mathcal{X}|\mathcal{Y} = y)$, which are the ancestors of $\mathcal{X}$ and $\mathcal{Y}$.

6. Reading

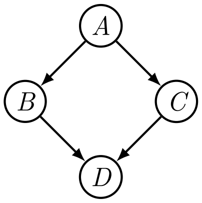
No recommended reading this week.

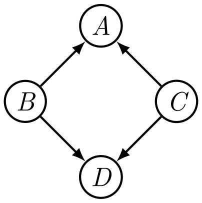
7. Exercises

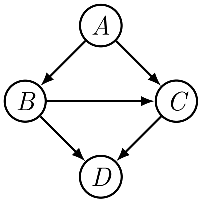
Exercise 1. *Prove that the following statements are equivalent for two nodes X and Y in an acyclic directed graph G:*

(i) X and Y are adjacent.
(ii) there is no set $\mathcal{Z}$ that d-separates X and Y.
(iii) X and Y are not *d-separated by an$(X) \cup$ an(Y).*
(iv) X and Y are not *d-separated by pa$(X) \cup$ pa(Y).*

Exercise 2. *Suppose you have an* indepedency oracle *I which answers independency queries "is $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$?". Discuss how you can use the result in the previous exercise and the oracle to decide which of the graphs below represent the independency relation induced by the oracle. Make an effort to find the most computationally efficient way, that is, one which calls the oracle the least number of times.*







Exercise 3. *Prove the following weaker version of the factorization theorem of Markov networks: If $p(\mathcal{V})$ is a joint distribution that factorizes as*

$$p(\mathcal{V}) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j(\mathcal{C}_j)\,,$$

where $\mathcal{C}_j$ are the maximal cliques of a graph G, then (G, p) satisfies all the local Markov properties. *Hint: use Proposition 1.*

8. Assignment

- Write code that receives a Markov network in UAI format and computes the partition function.
- Write a report describing in detail your implementation, with use cases.
- Solve the exercises.

Submit your report electronically via PACA as a single pdf file along with your source code. The exercises should be submitted also electronically via PACA *in a different pdf file.* Your solution should be elegant: correct, as simple as possible and clear.

The UAI File format can be obtained at `http://www.cs.huji.ac.il/project/PASCAL/fileFormat.php`.

References

[1] P.P. Shenoy. Valuation networks and conditional independence. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 191–199, 1993.

[2] F.G. Cozman. Credal networks. *Artificial Intelligence* 120, pp. 199–233, 2000.

[3] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artifical Intelligence*, pp. 220–227, 1990.