Introduction to Undirected Models – Markov Random Fields

Prof. Nicholas Zabaras
School of Engineering
University of Warwick
Coventry CV4 7AL
United Kingdom

Email: nzabaras@gmail.com
URL: http://www.zabaras.com/

September 2, 2014
Contents


- Factor Graphs, Factor Graphs from Undirected Graphs, Factor Graphs from Directed Graphs, Factor Graphs from Polytrees, Conditional Independence and Factorization

- Problems with Undirected Graphs and Factor Graphs, I-Map, D-Map and Perfect Map, Venn Diagram
  - Kevin Murphy, Machine Learning: A probabilistic Perspective, Chapter 19
  - Chris Bishop, Pattern Recognition and Machine Learning, Chapter 8
  - Video Lectures on Machine Learning, Z. Gahramani, C. Bishop and others.
Inference problems in Bayesian networks are often solved by turning the directed graph into another type of graphical model: Markov Random Fields or Factor Graphs.

The resulting inference algorithms to be introduced later on (e.g. Belief propagation) apply to all graphical models regardless of the starting representation.

Undirected graphs can model symmetric (non-causal) interactions that directed models cannot.
Markov Random Fields (MRFs) are undirected graphical models where nodes correspond to variables and undirected edges indicate independence. The parent/child asymmetry is removed as well as the subtleties related with head-to-head nodes.

In undirected graphical models, the graph semantics are much easier: $X$ and $Y$ are conditionally independent given $S$ if they are separated in the graph by $S$ (i.e. all paths from $X$ to $Y$ go via $S$).

In our example, one can easily see (we will revisit the Bayes Ball algorithm for undirected graphs) how the two independence relations are satisfied.

For a set of $M$ distinct random variables, each of which could have a link to any of the other $M-1$ nodes, making a total of $M(M - 1)$ links. Since each link is counted twice, so totally we have $2^{M(M-1)/2}$ distinct undirected graphs (each link is on or off).

The set of $8 = 2^{3(3-1)/2}$ possible graphs over three nodes is shown in Figure below.
Conditional Independence and Reachability

- $X_A$ is independent of $X_C$ given $X_B$ if the set of nodes $X_B$ separates the nodes $X_A$ from the nodes $X_C$, where “separation” means naive graph-theoretic separation.

- If every path from a node in $X_A$ to a node in $X_C$ includes at least one node in $X_B$, then $X_A \perp X_C \mid X_B$ holds.

- “$X_A \perp X_C \mid X_B$ holds for a graph $G$” implies that every member of the family of probability distributions associated with $G$ exhibits that conditional independence.

- “$X_A \perp X_C \mid X_B$ does not hold for a graph $G$” means that some distributions in the family associated with $G$ do not exhibit that conditional independence.
To answer conditional independence queries for undirected graphs, we remove $X_B$ from the graph and ask whether there are any paths from $X_A$ to $X_C$.

This is a "reachability" problem in graph theory and standard search algorithms provide a solution.
Consider two nodes \( i \) and \( j \) in the graph not connected via an edge. Then graph separation leads to the following conditional independence relation:

\[
p(x_i, x_j \mid x_{\{i,j\}}) = p(x_i \mid x_{\{i,j\}}) p(x_j \mid x_{\{i,j\}})
\]

The above CI relation leads us to consider the joint probability distribution as a product of potentials each defined over a clique, i.e. a fully connected set of nodes (a subset of nodes of the graph such that there exists a link between ALL pair of nodes in the subset).
Why We Need Other Types of Networks?

- We cannot find a Bayesian network that encodes ONLY the following two independence relations:

  \[ A \perp B \mid C \cup D \]
  \[ C \perp D \mid A \cup B \]

- For producing a graphical network with these properties, we need to introduce undirected graphical networks.
Markov Random Fields as graphical models are complementary to directed graphs (Bayesian networks).

Directed vs. Undirected Graphs

Cannot find an undirected perfect graph with 3 variables with the same property

Cannot find a directed perfect graph with these relations

Together they provide modeling power.
Clique: a subset of the nodes in a graph such that all the pairs of nodes are connected (a set of fully connected nodes)

Maximal clique: a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

In this example, if you connect all nodes, there is no link between nodes $x_1$ and $x_4$. 
The joint distribution for undirected graphs is written as a product of non-negative functions over the cliques of the graph:

\[ p(X) = \frac{1}{Z} \prod_c \psi_c(x_c) \]

where \( \psi_c(x_c) \) are the **clique potentials** and \( Z \) is a normalization constant.

Note the variables appearing in each clique come now in a symmetric form.

A distribution \( p \) that is represented by a UG \( H \) in this way is called a **Gibbs distribution over \( H \)**.
For the graph shown here, the partition of \( p(w,x,y,z) \) is:

\[
p(w,x,y,z) = \frac{1}{Z} \Psi_A(w,x,y) \Psi_B(x,y,z)
\]

where \( Z \) is a normalization constant, \( \Psi_A \) and \( \Psi_B \) are arbitrary functions corresponding to clique A and clique B, respectively.

Here, we have 2 maximal cliques:

- Clique A: \( \psi_A(W,X,Y) \)
- Clique B: \( \psi_B(X,Y,Z) \)
Consider the chain model shown below:

\[ X \quad Y \quad Z \]

Graph separation implies that: \( X \perp Z \mid Y \)

Using this, we can factor the joint \( p(x,y,z) \) as follows:

\[
p(x, y, z) = p(y) p(x \mid y) p(z \mid x, y) = p(y) p(x \mid y) p(z \mid y) \Rightarrow \\
p(x, y, z) = p(y) p(x \mid y) p(z \mid y) = p(x, y) p(z \mid y) \\
p(x, y, z) = p(y) p(x \mid y) p(z \mid y) = p(x \mid y) p(y, z) = p(x \mid y) p(y, z) \\
p(x, y, z) = p(y) p(x \mid y) p(z \mid y) = p(x \mid y) p(y, z)
\]

We cannot have all potentials as marginals or conditionals!

The positive clique potentials can only be thought as `compatibility' or `happiness' functions over their variables but not as probability distributions.
The simple graph separation criterion places constraints on the distributions associated with the undirected graph. This is clarified by the Hammersley-Clifford theorem.

**Hammersley-Clifford Theorem**: Any distribution that is consistent with an undirected graph has to factor according to the maximal cliques in the graph

\[
p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)
\]

where \(x_c\) are the variables in clique \(c\). The clique potential \(\psi(x_c)\) is any positive valued functions of the variables in clique \(c\).

Maximal cliques are shown with dotted lines.

- [Hammersley, J. M.; Clifford, P. (1971), *Markov fields on finite graphs and lattices*]
Let us define two families of distributions.

$\mathcal{U}_1$ is the family of distributions (parametric description of joint probability distributions) ranging over all possible choices of positive potential functions on the maximal cliques of the graph.

\[ p(x_1, ..., x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c), \quad Z = \sum_{x_1, ..., x_n} \prod_{c \in C} \psi_c(x_c) \]

$\mathcal{U}_2$ is the family of distributions $p(x_1, ..., x_n)$ that satisfies all conditional independence relations associated with the graph $G$.

Hammersley-Clifford Theorem: It states that $\mathcal{U}_1 = \mathcal{U}_2$.

- Hammersley, J. M.; Clifford, P. (1971), *Markov fields on finite graphs and lattices*
Suppose we model with finite elements the temperature in a house and let the temperature in each room is represented with a node in a graph with the corresponding random variable $x_i$ at node (room) $i$ taking the values +1 (hot) or -1 (cold).

The temperatures at neighboring rooms (nodes) depend on each other.

We thus couple the temperatures through real valued potential functions $\theta_{ij}(x_i, x_j)$. This will lead to a representation of the joint probability distribution.
The joint probability distribution of all variables in the graph that results from the coupling of the random variables $x_i$ and $x_j$ at the edge $(i,j)$ takes the form:

$$P(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{Z(\theta)} \sum_{(i,j) \in E} e^{\theta_{ij}(x_i, x_j)}$$

where the normalization factor is defined as:

$$Z(\theta) = \sum_{x_1, x_2, \ldots, x_n} \sum_{(i,j) \in E} e^{\theta_{ij}(x_i, x_j)}$$

An Example of MRF
The maximal cliques correspond to edges of the graph:

\[ P(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{Z(\theta)} e^{\sum_{(i,j) \in E} \theta_{ij}(x_i, x_j)} = \frac{1}{Z(\theta)} \prod_{(i,j) \in E} e^{\theta_{ij}(x_i, x_j)} \]

The exponential form of the potentials automatically enforces the positivity of the potentials (leading to a Boltzmann like joint distribution).
This joint probability distribution is consistent with all the independence relations implied by the undirected graph and the simple graph separation criterion.

\[ P(x_1, x_2, \ldots x_n ; \theta) = \frac{1}{Z(\theta)} \sum_{(i,j) \in E} \theta_{ij}(x_i, x_j) \]

\[ X \perp Y | S \]

Undirected Graphs and Independence Relations
Consider a specification

\[
p(x) = \frac{1}{Z} \prod_{C} g_j(x_{C_j}), \quad x = (x_1, x_2, ..., x_K), \quad C_j \subseteq \{1, 2, ..., K\}, \quad x_S \equiv (x_k : k \in S)
\]

How do we specify the graph based on this factorization?

- Create a node for each variable and connect any nodes \( i \) and \( k \) if there exists a set \( i \in C_j \) and \( k \in C_j \).

- These sets are the cliques of the graph (fully connected subgraphs).
Consider a specification

$$p(x) = \frac{1}{Z} \prod_{C} g_j(x_{C_j}), x = (x_1, x_2, \ldots, x_K), C_j \subseteq \{1, 2, \ldots, K\}, x_S \equiv (x_k : k \in S)$$

A clique is a fully connected subgraph. By clique we usually mean maximal clique (i.e. not contained within another clique)

Associated with each clique $C_i$ is the non-negative function $g_i$ which measures compatibility between settings of the variables.

Let $C_1 = \{A, C\}$, $A \in \{0, 1\}, C \in \{0, 1\}$

What does this mean?

<table>
<thead>
<tr>
<th>$A$</th>
<th>$C$</th>
<th>$g_1(A, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.2</td>
</tr>
</tbody>
</table>
The problem of determining CI is a reachability problem in classical graph theory: Starting from $X_A$ if you could not reach $X_B$ through any path in the graph, then $X_A$ and $X_B$ are conditionally independent.

The problem can be solved with standard search algorithms.

The two main rules when applying Bayes ball algorithm are shown above. A shaded node indicates that the network flow is blocked at that node.
For the graph below, using the two earlier Bayes ball rules, one can show that:

$$X_1 \perp X_5 \mid \{X_2, X_4\}$$
Conditional Independence and Factorization

- Conditional independence given by simple graph separation

\[ x \perp z \mid y \]

\[ p(x, z \mid y) = p(x \mid y)p(z \mid y) \]

- By specifying the variables \( y \), you block every path from the variables \( x \) to the variables \( z \) - making them independent.

- Every distribution that factorizes with this graph needs to satisfy the above conditional independence relation.
Conditional Independence and Factorization

- Conditional independence given by graph separation

\[ x \perp z \mid y \]

\[ p(x, z \mid y) = p(x \mid y)p(z \mid y) \]

- Consider all possible paths that connect nodes \( x \) to nodes \( y \). If all such paths pass through one or more nodes in set \( y \), then all such paths are 'blocked' and we have \( x \perp z \mid y \).

- Note that the nodes \( x \) and \( z \) cannot belong on the same clique. Thus the joint distribution factorizes as:

\[ p(...) = \frac{1}{Z} \Psi_A(x,..)\Psi_B(z,..) \]

\[ x \text{ and } z \text{ do not appear on the same } \Psi \]

- This highlights the proof of the CI of \( x \) & \( z \).
Conditional Independence and Factorization

- Conditional independence given by graph separation

- If the graph was fully connected, then the whole graph is a clique.

- Then any >0 function of all variables can be represented with this graph.
The maximal cliques in this graph are \{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_4\}, \{X_3, X_5\}, and \{X_2, X_5, X_6\}. For binary nodes, we represent the joint distribution on the graph via the potential tables.

\[
p(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{1}{Z} \psi(x_1, x_2, x_2, x_4, x_3, x_5, x_2, x_5, x_6)
\]
We cannot in general represent the potentials in maximal cliques with the corresponding marginal distributions.

![Graph with nodes X, Y, Z and edge (X ⊥ Z | Y)]

\[ p(x, y, z) = \frac{p(y)p(x \mid y)p(z \mid y)}{\Psi_A(x, y) \Psi_B(y, z)} \neq p(x, y)p(y, z) \]

In general, potential functions do not have a local probabilistic interpretation.

Potential functions often interpreted in terms of “energy”.

A potential function favors certain local configurations of variables. The global configurations that have high probability are those that satisfy as many of the favored local configurations as possible.
Consider a 1D spin model, \( X_i \in \{-1, 1\}, i = 0, \ldots, n \). If \( X_i = 1 \), then its neighbors \( X_{i-1} \) and \( X_{i+1} \) are likely to be spin up as well (and the opposite). This can be encoded with the Tables of the potential functions shown below:

We represent the potentials as:

\[
\psi_{X_c}(x_C) = e^{-H_{X_c}(x_C)} \Rightarrow p(x) = \frac{1}{Z} \prod_{C \in e} e^{-H_{X_c}(x_C)} = \frac{1}{Z} e^{-\sum_{C \in e} H_{X_c}(x_C)} = \frac{1}{Z} e^{-H(x)}
\]

energy: \( H(x) = \sum_{C \in e} H_{X_c}(x_C) \)
Without a topological ordering, there is no natural way to express the joint as a product of consistent local conditional or marginal probabilities; have to sacrifice local normalization.

A consequence is that some parts of the model may end up carrying more “weight” than others.

This can actually be useful, e.g., in discriminative classification.

However, it makes interpretation rather difficult.
Normalization Coefficient \( Z \)

Let us denote a clique by \( C \) and the set of variables in that clique by \( x_C \). Then the joint distribution is

\[
p(x) = \frac{1}{Z} \prod_C \psi_c(x_C)
\]

where \( \psi_c(x_C) \) is the potential over clique \( C \) and

\[
Z = \sum_x \prod_C \psi_c(x_C)
\]

is the normalization coefficient; note: \( M \) K-state variables \( \rightarrow K^M \) terms to sum in \( Z \).

Energies and the Boltzmann distribution

\[
\psi_c(x_C) = \exp\{-E(x_C)\}
\]
There are CI relations unique to directed or undirected graphs.

- **X ⊥ Z**
- **X ± Z | Y**

We can see from the two undirected graphs above, that there is no way we can capture the CI statements in the explaining away directed graph with three nodes.

No CI relations
Consider the four node undirected graphical model below.

\[ X_1 \perp X_4 \mid \{X_2, X_3\} \quad X_2 \perp X_3 \mid \{X_1, X_4\} \]

The graph above can e.g. represent a disease model with \( X_1, X_4 = \text{males}, X_2, X_3 = \text{female} \) and looking at diseases that can be transmitted only between opposite sex partners.

There is no directed graph with four nodes that can represent these CI relations.
What is the minimum set of nodes that will make node $i$ independent of the rest of the graph?

This is the set of neighbors. Given the neighbors of $X$, the variable $X$ is conditionally independent of all other variables:

$$X \perp Y \mid ne(X), \forall Y \notin \{X \cup ne(X)\}$$

Given the neighbors of $X$, $X$ is conditionally independent of all other variables.

$V$ is a Markov Blanket for $X$ iff $X \perp Y \mid V, \forall Y \notin \{X \cup V\}$. Markov boundary is the minimal Markov Blanket which is the $ne(X)$ for undirected graphs.
The Markov blanket for an undirected graph takes a particularly simple form, because a node will be conditionally independent of all other nodes conditioned only on the neighboring nodes (it protects the node from the rest of the variables).
Knowing the temperature at some nodes (rooms) \( i \) and \( j \), what can we say about the temperature at other nodes \( k \)?

We answer these questions by computing the posterior marginals as shown below:

\[
P(x_k \mid x_i, x_j; \theta) = \frac{P(x_k, x_i, x_j; \theta)}{\sum_{x_k \in \{-1, 1\}} P(x_k, x_i, x_j; \theta)}
\]

where (using a trick to enforce the observed data):

\[
P(x_k, x_i, x_j; \theta) = \sum_{x_1, \ldots, x_n \setminus x_k} \delta(x_i, x_i) \delta(x_j, x_j) P(x_1, x_2, \ldots, x_n; \theta)
\]
Computing the posterior marginals is easier if the model had a tree structure (unique path of influence between any two nodes).

We will thus need to first review inference on trees before generalizing on other graphs.
We see that the factor \( p(x_4|x_1, x_2, x_3) \) involves the four variables \( x_1, x_2, x_3, \) and \( x_4, \) and so these must all belong to a single clique.

To ensure this, we add extra links between all pairs of parents of the node \( x_4. \)

This process is called moralization, and the resulting undirected graph is called the moral graph.

Note: the moral graph in this example exhibits no conditional independence properties (in general moralization adds the fewest extra links and so retains the max number of independence properties).

Note: Moralization & conversion of directed to undirected graphs is important in the junction tree algorithm (Exact Inference).
Converting Directed to Undirected Graphs

\[
p(x) = p(x_1) \underbrace{p(x_2 | x_1)}_{\psi_{12}(x_1, x_2)} \underbrace{p(x_3) p(x_4 | x_2, x_3)}_{\psi_{234}(x_2, x_3, x_4)} \quad p(x) = \frac{1}{Z} \psi_{12}(x_1, x_2) \psi_{234}(x_2, x_3, x_4)
\]

- **Step 1**: Moralize (marry the co-parents) and omit edge directions.

- **Step 2**: Map each conditional probability to a clique potential that contains it (mapping is not necessarily unique). \(Z=1\) in this case.
Converting Directed to Undirected Graphs

\[
p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = p(x_3) p(x_1 | x_3) p(x_2 | x_3) p(x_4 | x_3) p(x_5 | x_4) p(x_6 | x_4) p(x_7 | x_4) = \\
= p(x_3) \frac{p(x_1, x_3)}{p(x_3)} \frac{p(x_2, x_3)}{p(x_3)} \frac{p(x_3, x_4)}{p(x_3)} \frac{p(x_4, x_5)}{p(x_4)} \frac{p(x_4, x_6)}{p(x_4)} \frac{p(x_4, x_7)}{p(x_4)} = \\
= \frac{\text{product of cliques}}{\text{product of clique intersections}} = \\
= g_1(x_1, x_3) g_2(x_2, x_3) g_3(x_3, x_4) g_4(x_4, x_5) g_5(x_4, x_6) g_6(x_4, x_7) = \\
= \prod_i g_i(C_i)
\]

This way we can convert any directed tree to an undirected tree with the same independence relations.
Converting Directed to Undirected Graphs

- Consider first the chain like graph where there is direct correspondence between the conditional probabilities and the corresponding potentials.
- Conversion of the undirected graph to a directed graph can also work similarly by selecting node 1 as the root node and pointing towards the leaf nodes.

\[
p(x) = p(x_1) p(x_2 | x_1) p(x_3 | x_2) \cdots p(x_N | x_{N-1})
\]

\[
p(x) = \frac{1}{Z} \psi_{1,2}(x_1, x_2) \psi_{1,2}(x_2, x_3) \cdots \psi_{N-1,N}(x_{N-1}, x_N)
\]
From Directed to Undirected Trees

- Undirected Tree: there is only one path between any pair of nodes. Such graphs have no loops.

- Directed Tree: there is a single root node, all the other nodes have only one parent.

Converting a directed tree to an undirected tree is simple by taking the two-node potentials as $p(x_k|pa_k)$. 
From Undirected to Directed Trees

- A parent can have many children and each of the conditionals defines a separate potential term. $p(x_1)$ can be represented either as a single node potential or incorporated into the potential associated with the root node.

- To convert an undirected tree to a directed tree, you simply pick a root node (node 4 in the Fig) and direct all edges pointing away towards the leaf nodes. By normalizing the potential of each edge, we obtain the corresponding conditionals in the directed graph.

- From an undirected graph, we can produce $N$-directed graphs (one for each selection of the root node).
Summary of Factorization Properties

- For directed graphs

\[ p(x_1, x_2, ..., x_D) = \prod_{i=1}^{D} p(x_i | pa_i) \]

and conditional independence comes from d-separation.

- For undirected graphs:

\[ p(x) = \frac{1}{Z} \prod_c \psi_c(x_c) \]

and conditional independence comes from graph separation.
This is a typical application of undirected graphs in image denoising (we observe $y_i$ and we want to compute the labels $x_i \in \{0,1\}$).

We need the posterior of $x$.

Note that here there are loops (so you need approximations).

There are two types of cliques in the model, $\{x_i, x_j\}$ and $\{x_i, y_i\}$.

\[
p(x, y) = \frac{1}{Z} \prod_i \Phi_i(y_i, x_i) = \frac{1}{Z} \prod_{i,j} \Psi_i(x_i, x_j) \prod_i \Xi_i(x_i, y_i)
\]
The Markov random field model is shown in the figure.

There are two types of cliques in the model, \( \{x_i, x_j\} \) and \( \{x_i, y_i\} \).

We also add an extra term \( h x_i \) for each pixel \( i \), in order to bias the model towards pixel values that have one particular sign in preference to the other (this is like multiplying with an additional potential from a sub-clique of the maximal clique \( \{x_i, x_j\} \)).

The energy function:

\[
E(x, y) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i, \quad \beta, \eta > 0
\]

The joint distribution:

\[
p(x, y) = \frac{1}{Z} \exp\{ -E(x, y) \}
\]
Conditional Markov Random Field

- Let's look at the conditional distribution of the $x$'s for given $y$.

$$ p(x | y) = \frac{1}{Z(y)} \prod_{i,j} \Psi_i(x_i, x_j; y) \prod_i \Xi_i(x_i; y) $$

- Here $y$ is a high-dimensional vector but the labels $x$ are low-dimensional.

- Note that in this model the potentials $\Psi_i(x_i, x_j; y)$ depend on all $y$'s - in the joint distribution model there was no $y$ dependence!

- Also note that in the model of the joint distribution, each $x_i$ depends only on the $y_i$ at that pixel. This is not the case in the conditional MRF here.
Illustration: Image De-Noising

Original Image

Noisy Image: Randomly Changing 10% of the pixels of the image on the left

Iterative Conditional Modes (ICM)

The idea for the Iterative conditional modes method is to find the image with minimum energy/maximum probability:

- first to initialize the variables \{x_i\}, which we do by simply setting \(x_i = y_i\) for all \(i\).

- take one node \(x_j\) at a time and we evaluate the total energy for the two possible states \(x_j = +1\) and \(x_j = -1\), keeping all other node variables fixed, and set \(x_j\) to whichever state has the lower energy.

- repeat the update for another site, and so on, until some suitable stopping criterion is satisfied (converge).

Illustration: Image De-Noising

Restored Image (ICM)

$\beta=1, \eta=2.1, h=0$

96% of the pixels agree with the original image (ICM only finds a local maximum)

Restored Image (Graph cuts)

99% of the restored pixels agree with the original image (Graph Cuts locate the global maximum for this problem)

Consider an undirected graphical model $G$ with nodes $X_1, ..., X_n$ and strictly positive potential functions.

A subset of all distributions, $UI \subseteq U$, maintain the CI assertions implied by graph separation in $G$.

Another subset of distributions, $UF \subseteq U$, can be factored according to the maximal cliques of $G$.

The theorem establishes that $UI = UF$.

- Hammersley, J. M.; Clifford, P. (1971), *Markov fields on finite graphs and lattices*
Hammersley-Clifford

- Helped establish that global distributions that emerge from local interactions could be characterized and analyzed.

- Influential in many areas of statistics, including:
  - Geographical epidemiology
  - Image analysis
  - Analysis of contingency tables (log-linear models)

- Intimately connected with Markov chain Monte Carlo methods, statistical mechanics

  - Hammersley, J. M.; Clifford, P. (1971), *Markov fields on finite graphs and lattices*
Factor graphs explicate how the joint distribution factors into smaller components.

Each factor node is connected to all the variable nodes that the corresponding factor depends on.

\[ p(x) = p(x_1) p(x_2) p(x_3 | x_1, x_2) \]

The conversion of a directed graph to a factor graph is illustrated in the Figure below.

\[
p(x) = p(x_1)p(x_2)
\]

\[
f(x_1, x_2, x_3) = p(x_3 | x_1, x_2)p(x_1)p(x_2)
\]

Again, there can be multiple factor graphs all of which correspond to the same directed graph.
We can also define a factor graph representation of an undirected graph.

Each factor node is connected to all the variable nodes that the corresponding factor depends on.
An undirected graph can be readily convert it to a factor graph.

\[ \psi(x_1, x_2, x_3) \]

\[ f(x_1, x_2, x_3) = \psi(x_1, x_2, x_3) \]

Note that there may be several different factor graphs that correspond to the same undirected graph.
All nodes in (a), (b), and (c) have exactly the same neighbors and these three graphs represent exactly the same conditional independence relationships.

In (c) the probability factors into a product of pairwise functions.

Consider the case where each variable is discrete and can take on $K$ possible values. The functions in (a) and (b) are tables with $O(K^3)$ cells, whereas in (c) they are $O(K^2)$. 
In the case of a directed polytree, conversion to an undirected graph results in loops due to the moralization step, whereas conversion to a factor graph again results in a tree.
Let us write the joint distribution over a set of variables in the form of a product of factors

\[ p(x) = \prod_s f_s(x_s) \]

For example, a distribution below can be expressed as a factor graph shown in the figure.

\[ p(x) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3) \]

- **Circles**: random variables
- **Filled dots**: Factors in the joint distribution
- **Neighbors**: Two nodes are neighbors if they share a factor
Factor Graphs

The $g_i$ are non-negative functions of their arguments, and $Z$ is a normalization constant e.g. in the fig. on the left, if all variables are discrete and take values in $A \times B \times C \times D \times E$, then

$$P(A,B,C,D,E) = \frac{1}{Z} g_1(A,C)g_2(B,C,D)g_3(C,D,E)$$

$$P(A,B,C,D,E) = \frac{1}{Z} g_1(A,C)g_2(B,C)g_3(C,D)g_4(B,D)g_5(C,E)g_6(D,E)$$

$$Z = \sum_{a \in A} \sum_{b \in B} \sum_{c \in C} \sum_{d \in D} \sum_{e \in E} g_1(A = a, C = c)g_2(B = b, C = c, D = d)g_3(C = c, D = d, E = e)$$
Factor Graphs and CI Relations

- A path is a sequence of neighboring nodes.
- \( X \perp Y \mid \forall V \) if every path between \( X \) and \( Y \) contains some node \( V \in \mathcal{V} \)
- Given the neighbors of \( X \), the variable \( X \) is conditionally independent of all other variables (same as in undirected graphs):
  \[
  X \perp Y \mid ne(X), \forall Y \notin \{X \cup ne(X)\}
  \]

Every path from \( X \) to \( Y \) has to go through its neighbors.
Conditional Independence and Factorization

- Let's consider the following conditional independence:

\[ X \perp Y | V \iff p(X | Y, V) = p(X | V) \]

- This independence relation is represented with the factorization:

\[ P(X, Y, V) = \frac{1}{Z} g_1(X, V) g_2(Y, V) \]

- Indeed:

\[ P(Y, V) = \sum_x P(X, Y, V) = \frac{1}{Z} \sum_x g_1(X, V) g_2(Y, V) \]

and

\[ P(X | Y, V) = \frac{P(X, Y, V)}{P(Y, V)} = \frac{\frac{1}{Z} g_1(X, V) g_2(Y, V)}{\sum_x g_1(X, V) g_2(Y, V)} = \frac{g_1(X, V)}{\sum_x g_1(X, V)} \text{ (independent of } Y) \]

- Once more \textit{we go from factorization to independence relations}.
In UGs and FGs, many useful independencies are unrepresented—two variables are connected merely because some other variable depends on them.

This highlights the difference between marginal independence and conditional independence.

R and S are marginally independent (i.e. given nothing), but they are conditionally dependent given G. This relation cannot be represented with UG or FGs.

Also we have “Explaining Away”: Observing that the sprinkler is on, would explain away the observation that the ground was wet, making it less probable that it rained.
I-Map and Perfect Map

- **D map**: A graph is said to be a D map (for 'dependency map') of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.

  A completely disconnected graph (no links) will be a trivial D map for any distribution.

- **I map**: every conditional independence statement implied by a graph is satisfied by a specific distribution, then the graph is said to be an I map (for 'independence map') of that distribution.

  Clearly a fully connected graph will be a trivial I map for any distribution.

- **Perfect map**: every conditional independence property of the distribution is reflected in the graph, and vice versa.
Let \( P \) be the set of all distributions over a set of variables. The Venn diagram consists:

- the set of distributions such that for each distribution there exists a directed graph that is a perfect map (D).
- the set of distributions such that for each distribution there exists an undirected graph that is a perfect map (U).
- Other distributions (chain graphs) for which neither directed nor undirected graphs offer a perfect map.

Chain graphs represent are perfect maps for distributions broader than those corresponding to either directed or undirected graphs. There are of course distributions that even chain graphs cannot provide a perfect map.