Lecture 1

Graphical Models

Prof. Dahua Lin
dhlin@ie.cuhk.edu.hk

Roadmap

Graphical Models

- The key idea behind *graphical models* is **factorization**.
- A **graphical model** generally refers to a **family** of joint distributions over multiple variables that **factorize** according to the structure of the underlying graph.

Graphical Models

A **graphical model** can be viewed in two ways:

- A data structure that provides the skeleton for representing a joint distribution in a **factorized** manner.
- A compact representation of a set of **conditional independencies** about a family of distributions.

These two views are *equivalent* in a strict sense.

Distributions on a Graph

Consider a graph $G = (V, E)$, where edges can be *directed* or *undirected*.

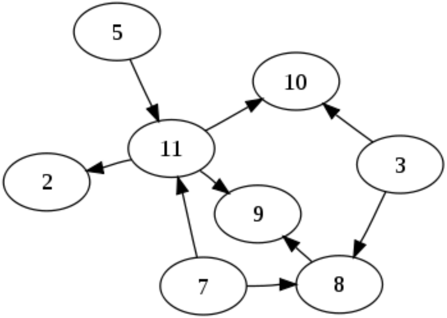
- Attach a random variable X_s to each vertex $s \in V$.
- The state space for X_s is denoted by $\mathcal{X}_s$.
- A particular instance of X_s is denoted by x_s.
- We can also consider a set of variables: X_A and x_A.

Categories of Graphical Models

- Bayesian Networks (Directed Acyclic Graphs)
- Markov Random Fields (Undirected Graphs)
- Chain Graphs (Directed acyclic graphs over undirected components)
- Factor Graphs

Directed Acyclic Graph

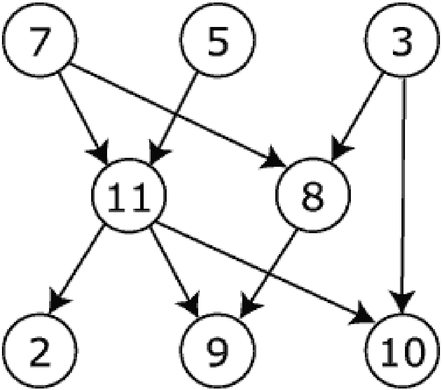
- Consider a *directed graph* $G = (V, E)$, G is called a **directed acyclic graph (DAG)** if it has no *directed cycles*.
- Given an edge $(s, t) \in E$, s is called a **parent** of t, and t is called a **child** of s.
- A vertex s is called an **ancestor** of t and t an **descendant** of s, denoted as $s \prec t$, if there exists a directed path from s to t.



Topological Ordering

- A **topological ordering** of a *directed graph* $G = (V, E)$ is a *linear ordering* of vertices such that for each edge $(s, t) \in E$, s always comes before t.

- A *finite directed graph* is *acyclic* if and only if it has a *topological ordering*.



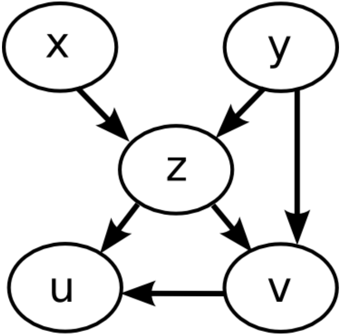
Bayesian Networks

Given a DAG $G = (V, E)$, we say a joint distribution over X_V **factorizes** according to G, if its density p can be expressed as:

$$p(x_V) = \prod_{s \in V} p_s(x_s | x_{\pi(s)})$$

- Such a model is called a **Bayesian Network** over G.
- $\pi(s)$ is the set of s's parents, which can be empty.

Bayesian Networks: Example



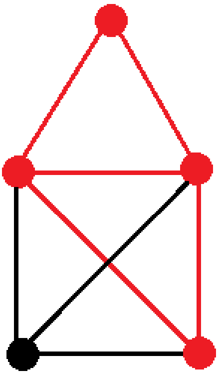
$$p(x) \cdot p(y) \cdot p(z|x, y) \cdot p(v|y, z) \cdot p(u|z, v)$$

Undirected Graphs and Cliques

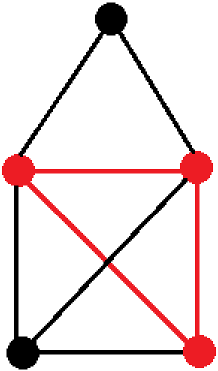
Consider an *undirected graph* $G = (V, E)$

- A **clique** is a fully connected subset of vertices
- A *clique* is called **maximal** if it is not **properly** contained in another *clique*.
- $\mathcal{C}(G)$ denotes the set of all **maximal cliques**.

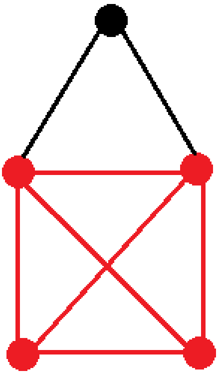
Undirected Graphs and Cliques (cont'd)



not a clique



non-maximal clique



maximal clique

Markov Random Fields

Consider an *undirected graph* $G = (V, E)$, we say a joint distribution of X_V **factorizes** according to G if its density p can be expressed as:

$$p(x_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

- This is called a *Markov Random Field* over G.
- $\psi_C : \mathcal{X}_C \to \mathbb{R}_+$ are called *factors*.

Markov Random Fields (cont'd)

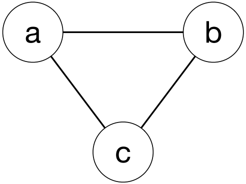
- The *normalizing constant* Z is usually needed to ensure the distribution is properly normalized:

$$Z = \int \prod_{C \in \mathcal{C}(G)} \psi_C(x_C)\mu(dx).$$

- Generally, the *compatibility functions* ψ_C need not have any obvious relations with the marginal or conditional distributions over the cliques.

MRF Parameterization

All MRFs can be parameterized in terms of *maximal cliques*. In practice, this is not necessarily the most natural way.



- Natural parameterization: $\frac{1}{Z}\psi_{ab}(x_a, x_b)\psi_{bc}(x_b, x_c)\psi_{ac}(x_a, x_c)$
- Maximal-clique based: $\frac{1}{Z}\psi'(x_a, x_b, x_c)$ with $\psi'(x_a, x_b, x_c) = \psi_{ab}(x_a, x_b)\psi_{bc}(x_b, x_c)\psi_{ac}(x_a, x_c)$

The graphical structure also encodes a set of conditional independencies among the variables.

Conditional Independence

- Consider a joint distribution over (X, Y, Z), X and Y are called *conditionally independent* given Z, denoted by $X \perp Y | Z$ iff

$$\Pr(X \in A \ \& \ Y \in B \mid Z) = \Pr(X \in A \mid Z)\Pr(Y \in B \mid Z) \ a.s.$$

More generally,

$$E_{X,Y|Z}[f(X)g(Y)] = E_{X|Z}[f(X)]E_{Y|Z}[g(Y)] \ a.s.$$

- Suppose the conditional distributions $X|Z$ and $Y|Z$ have densities $p_{X|z}$ and $p_{Y|z}$, then $X \perp Y \mid Z$, if the following equality holds *almost surely*:

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

I-map

- Let $\mathcal{P}$ be a family of distributions (*e.g.* a graphical model). We define $\mathcal{I}(\mathcal{P})$ to be the set of *conditional independencies* in the form of $(X \perp Y \mid Z)$ that hold for all distributions in $\mathcal{P}$.
- Given a graph G associated with a set of *conditional independencies* $\mathcal{I}(G)$, then G is called an *I-map* of $\mathcal{P}$ if $\mathcal{I}(G) \subset \mathcal{I}(\mathcal{P})$.
- An *I-map* is a graph that captures (part of) the conditional independencies of a distribution family.

Conditional Independencies of MRFs

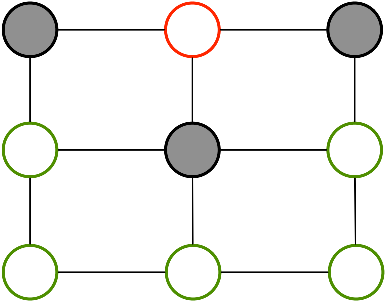
The conditional independencies of an MRF can be characterized in three ways:

- Local independencies
- Pairwise independencies
- Global independencies

In the sequel, we consider an *undirected graph* $G = (V, E)$.

Local Independencies

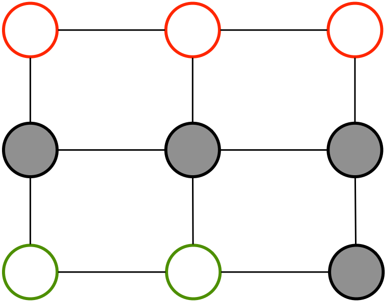
Local independencies: For each $s \in V$, X_s is *independent* of the rest given its neighbors $\mathcal{N}_G(s)$.



$$\mathcal{I}_l(G) = \left\{ X_s \perp X_{V \setminus (\{s\} \cup \mathcal{N}_G(s))} \mid X_{\mathcal{N}_G(s)} : s \in V \right\}$$

Pairwise Independencies

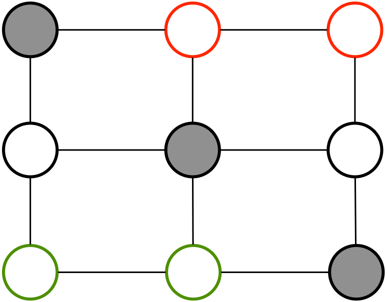
Pairwise independencies: Given two disjoint sets $A, B \subset V$ with no direct edges between them, X_A is *independent* of X_B given the rest.



$$\mathcal{I}_p(G) = \left\{ X_A \perp X_B \mid X_{V \setminus (A \cup B)} : A - B \notin G \right\}$$

Global Independencies

Global independencies: We say C *separates* A and B, denoted by $\mathrm{sep}(A, B \mid C)$, if all paths between A and B go through C. If C separates A and B, then X_A is *independent* of X_B given X_C.



$$\mathcal{I}_g(G) = \{X_A \perp X_B \mid X_C : \mathrm{sep}(A, B \mid C)\}$$

Relations between Independencies

- $\mathcal{I}_l(G) \subset \mathcal{I}_p(G) \subset \mathcal{I}_g(G)$
- Given a distribution or a family of distribution $\mathcal{P}$, we say $\mathcal{P}$ *satisfies* $\mathcal{I}$ if it satisfies all *conditional independencies* in $\mathcal{I}$, denoted by $\mathcal{P} \models \mathcal{I}$.
- Generally,

$$\mathcal{P} \models \mathcal{I}_g(G) \Rightarrow \mathcal{P} \models \mathcal{I}_p(G) \Rightarrow \mathcal{P} \models \mathcal{I}_l(G)$$

- If $\mathcal{P}$ is a family of *positive distributions*, then

$$\mathcal{P} \models \mathcal{I}_g(G) \Leftrightarrow \mathcal{P} \models \mathcal{I}_p(G) \Leftrightarrow \mathcal{P} \models \mathcal{I}_l(G)$$

Soundness

- Let P be a distribution that factorizes according to an undirected graph G, then $P \models \mathcal{I}(G)$, or in other words, G is an *I-map* of P.
- $P \models \mathcal{I}_p(G)$ and $P \models \mathcal{I}_l(G)$.
- How to prove?
  - How is the *separation assumption* related to the *maximal cliques*?

We have shown that if P factorizes according to G, then G is an I-map for P. Is the converse also true?

Hammersley-Clifford

- **Hammersley-Clifford Theorem:** Let P be a *positive distribution* over X_V and $G = (V, E)$ be an I-map of P, then P *factorizes* according to G.

- Combining *Soundness* and *Hammersley-Clifford*:

A *positive distribution* P *factorizes* according to G **if and only if** G is an *I-map* of P, *i.e.* $P \models \mathcal{I}(G)$.

Conditional Independencies of BN

Conditional independencies of a *Bayesian network* can be characterized in two ways:

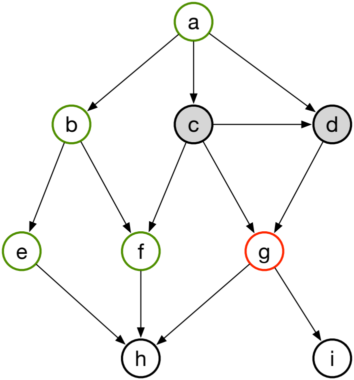
- Local independencies
- Global independencies (via d-*separation*)

In the sequel, we consider a *directed graph* $G = (V, E)$.

Local Independencies for BN

Given $s \in V$, X_s is independent of its *non-descendants* given its *parents*:

$$\left\{ X_s \perp X_{\mathrm{NonDesc}(s)} \mid X_{\pi(s)} : s \in V \right\}.$$



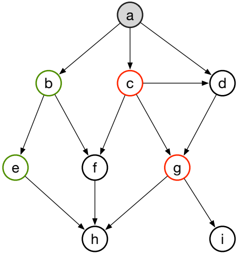
d-separation

- When *"influence"* can flow from X to Y via Z, we say that the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is **active**:
  - $X \to Z \to Y$ is *active* iff Z is not observed.
  - $X \leftarrow Z \leftarrow Y$ is *active* iff Z is not observed.
  - $X \leftarrow Z \to Y$ is *active* iff Z is not observed.
  - *(V-structure)* $X \to Z \leftarrow Y$ is *active* iff either Z or some of Z's descendants is observed.
- A trail $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ is called **active** when all sub-trails $X_{i-1} \rightleftharpoons X_i \rightleftharpoons X_{i+1}$ are *active*.
- Let A, B, C be three sets of vertices of G. A and B are **d-separated** by C, denoted by $\mathrm{dsep}(A, B \mid C)$, if there is neither direct link nor active trail between A and B when X_C are observed.

Global Independencies for BN

Given $A, B, C \subset V$, X_A is *independent* of X_B given X_C if A and B is d-*separated* by C on the graph G:

$$\mathcal{I}_g(G) = \{X_A \perp X_B \mid X_C : \mathrm{dsep}(A, B \mid C)\}$$

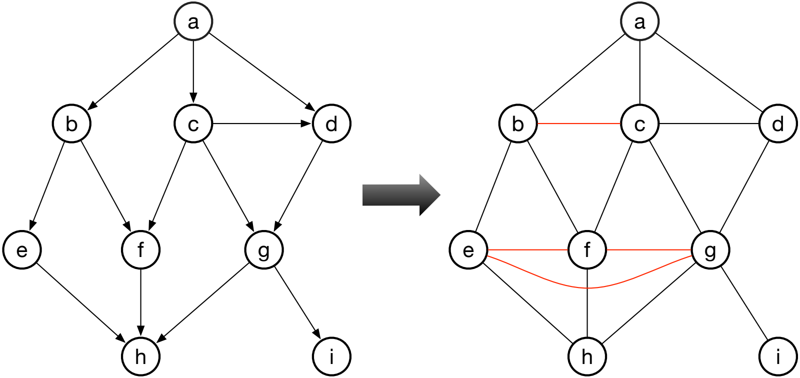


$\mathcal{I}_l(G) \subset \mathcal{I}_g(G)$. Also, if P *factorize* according to G, then $P \models \mathcal{I}_g(G)$, or we say G is an *I-map* of P, *i.e.* $P \models \mathcal{I}_g(G)$.

Moralized Graphs

- Given a *directed graph* $G = (V, E)$, we can construct a **moralized graph**, denoted by $\mathcal{M}[G]$ by adding edges between each node and its parents and between each node's parents.
- In $\mathcal{M}[G]$, the subgraph that spans $\{s\} \cup \pi(s)$ forms a *clique*, denoted by C_s.
- The procedure of constructing $\mathcal{M}[G]$ from G is called **moralization**.

Moralized Graphs (Illustration)



From BN to MRF

- If p factorizes according to G as

$$p(x_V) = \prod_{s \in V} p_s(x_s | x_{\pi(s)})$$

then p factorizes according to $\mathcal{M}[G]$:

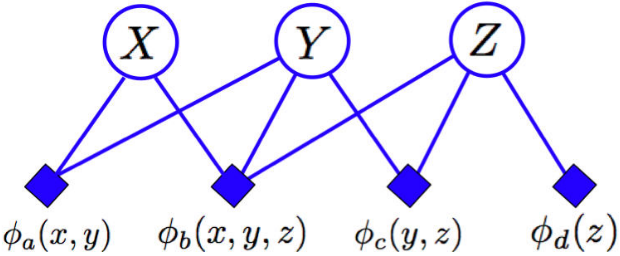
$$p(x_V) = \prod_{s \in V} \psi_s(x_{C_s}), \text{ with } \psi_s(x_{C_s}) = p_s(x_s | x_{\pi(s)})$$

- $\mathcal{I}(\mathcal{M}[G]) \subset \mathcal{I}(G)$. **Is the opposite true?**

Factor Graphs

- An MRF does not always *fully* reveal the *factorized structure* of a distribution.
- A *factor graph* can sometimes give a more accurate characterization of a family of distributions.
- A **factor graph** is a *bipartite graph* with links between two types of nodes: **variables** and **factors**.
- A variable x and a factor f is linked in a factor graph, if the factor involves x as an argument.

Factor Graphs (Illustration)



$$p(x,y,z) = \frac{1}{Z}\phi_a(x,y)\phi_b(x,y,z)\phi_c(y,z)\phi_d(z).$$