ECE 175B: Probabilistic Reasoning and Graphical Models
Lecture 10: Probabilistic Graphical Models and Their Properties

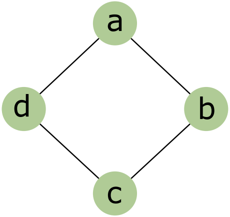
Florian Meyer & Ken Kreutz-Delgado

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

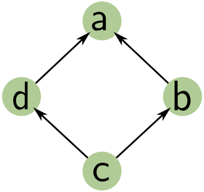
Limitations of BNs

- Consider there are four students a, b, c and d who are trying to clear a misunderstanding of a concept in class. We define their realizations as random variables $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d} \in \{0, 1\}$ where 0 means no misunderstanding and 1 means misunderstanding
- The students only interact in pairs and we know a, c never speak to each other directly and neither do b and d, i.e., we have the conditional independence statements $a \perp\!\!\!\perp c \mid b, d$ and $b \perp\!\!\!\perp d \mid a, c$
- Then the connection of them can be represented as a skeleton



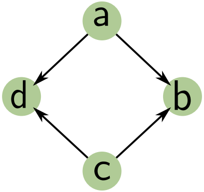
Limitations of BNs

- However, this cannot be captured by a BN, for example



$$a \perp\!\!\!\perp c \mid b, d$$

$$b \not\perp\!\!\!\perp d \mid a, c$$



$$b \perp\!\!\!\perp d \mid a, c$$

$$a \not\perp\!\!\!\perp c \mid b, d$$

Markov Networks (MNs) vs Bayesian Networks (BNs)

- We see that there are conditional independence statements that can't be captured by a BN
- Many, but not all, of these can be captured by a MN
- But we shall see that there are conditional independence statements that can't be captured by a MN yet can be captured by a BN
- The ability to encode conditional independence statements gives the "expressive power" of a graph
- The conditional independence statements give the "semantics" of the graph, i.e., they tell us what a graph "means"
- We are interested in understanding the relative expressive power of MNs and BNs
- Recall that a BN is a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that encodes a probability distribution factorization $P(\mathcal{X}) = \prod_{j=1}^{N} P(\mathrm{x}_j|\mathbf{pa}(\mathrm{x}_j))$
- So what about a MN?

Markov Network (MN)

- For a set of variables $\mathcal{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$, a Markov Network is a undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $j \leftrightarrow \mathrm{x}_j, j = 1, \ldots, N$, that encodes a distribution factorization

$$P(\mathcal{X}) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

  where $\mathcal{X}_c, c = 1, \ldots, C$ are cliques as a decomposition of $\mathcal{G}$; $\phi_c(\mathcal{X}_c) \geqslant 0, c = 1, \ldots, C$ are potential functions

- Z is a constant which ensures normalization, called the "partition function"

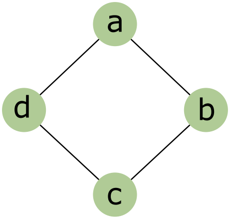
$$Z = \sum_{\mathrm{x} \in \mathcal{X}} \tilde{P}(\mathcal{X})$$

  where $\tilde{P}(\mathcal{X})$ is the unnormalized distribution as a product of all potentials, i.e.,

$$\tilde{P}(\mathcal{X}) = \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

Markov Network (MN)

- Consider the previous example, we have a MN as

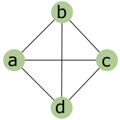


- Here, $\mathcal{X} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$ with $\{\mathcal{X}_c\}_{c=1}^{4} = \{\{\mathrm{a}, \mathrm{b}\}, \{\mathrm{b}, \mathrm{c}\}, \{\mathrm{c}, \mathrm{d}\}, \{\mathrm{d}, \mathrm{a}\}\}$
- The corresponding factorization of potentials is given as

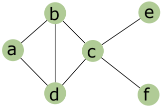
$$P(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}) = \frac{1}{Z}\phi_1(\mathrm{a}, \mathrm{b})\phi_2(\mathrm{b}, \mathrm{c})\phi_3(\mathrm{c}, \mathrm{d})\phi_4(\mathrm{d}, \mathrm{a})$$

Markov Network (MN)

- Some more examples



$$\begin{aligned}\tilde{P}(a,b,c,d) &= \overbrace{\varphi_1(a,b,c,d)}^{\text{maximal cliques}} \\ &= \underbrace{\phi_1(a,b,d)\phi_2(a,b,c)\phi_3(a,d,c)\phi_4(b,c,d)}_{\text{non-maximal cliques}}\end{aligned}$$



$$\begin{aligned}\tilde{P}(a,b,c,d,e,f) &= \overbrace{\varphi_1(a,b,d)\varphi_2(b,c,d)\varphi_3(c,e)\varphi_4(c,f)}^{\text{maximal cliques}} \\ &= \underbrace{\phi_1(a,b)\phi_2(a,d)\phi_3(b,d)}_{\text{non-maximal cliques}}\underbrace{\phi_4(b,c,d)\phi_5(c,e)\phi_6(c,f)}_{\text{maximal cliques}}\end{aligned}$$

- Note that the cliques from graph decomposition is not unique, we can choose any set of cliques if only the union of cliques covers the whole graph
- Different clique decomposition yields different factorization of potentials; some clique choices yield potential functions that are more interpretable
- In fact, we can consider functions of the maximal cliques, without loss of generality, because other cliques must be subsets of maximal cliques

Math Fact

- **Theorem:** $x \perp\!\!\!\perp y \mid z$, *i.e.,* $P(x, y|z) = P(x|z)P(y|z)$ *if and only if there exists two function* $f(x, z)$ *and* $g(y, z)$ *such that* $P(x, y|z) = f(x, z)$ $g(y, z)$ *over domain of* x, y, z

- **Proof:** "Only if" is trivial, just let $f(x, z) = P(x|z)$ and $g(y|z) = P(y|z)$
  Now we prove "if": Assume $P(x, y|z) = f(x, z)g(y, z)$, we have

$$1 = \sum_x \sum_y P(x, y|z) = \Big(\sum_x f(x, z)\Big)\Big(\sum_y g(y, z)\Big)$$

$$P(x|z) = \sum_y P(x, y|z) = f(x, z)\Big(\sum_y g(y, z)\Big)$$

$$P(y|z) = \sum_x P(x, y|z) = g(y, z)\Big(\sum_x f(x, z)\Big)$$

A Math Fact

- **Proof Cont'd:** Then we have

$$
\begin{aligned}
P(\mathrm{x}, \mathrm{y}|\mathrm{z}) &= f(\mathrm{x}, \mathrm{z}) \cdot 1 \cdot g(\mathrm{y}, \mathrm{z}) \\
&= \underbrace{f(\mathrm{x}, \mathrm{z}) \Big(\sum_{\mathrm{y}} g(\mathrm{y}, \mathrm{z}) \Big)}_{P(\mathrm{x}|\mathrm{z})} \underbrace{\Big(\sum_{\mathrm{x}} f(\mathrm{x}, \mathrm{z}) \Big) g(\mathrm{y}, \mathrm{z})}_{P(\mathrm{y}|\mathrm{z})} \\
&= P(\mathrm{x}|\mathrm{z}) P(\mathrm{y}|\mathrm{z})
\end{aligned}
$$

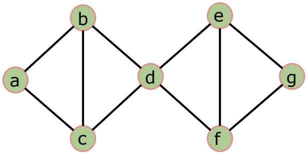
MN Graph Separation & The Global Markov Property

- **Markov Graph Separation:** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be disjoint node subsets of $\mathcal{V}$ in MN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we say that $\mathcal{Z}$ separates $\mathcal{X}$ and $\mathcal{Y}$, denoted as $<\mathcal{X}|\mathcal{Z}|\mathcal{Y}>_{\mathsf{d}}$ if and only if every path from $\mathcal{X}$ to $\mathcal{Y}$ passes through $\mathcal{Z}$

- **Global Markov Property:** For disjoint sets of variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, if $<\mathcal{X}|\mathcal{Z}|\mathcal{Y}>_{\mathsf{d}}$ in the corresponding MN, then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ (The proof is based on our "math fact")

Global Markov Property

- **Example:**



We have a set of random variables $\mathcal{X} = \{\mathrm{a}, \ldots, \mathrm{g}\}$

For $< \mathrm{a}|\mathrm{d}|\mathrm{g} >_{\mathrm{d}}$, we want to show that $\mathrm{a} \perp\!\!\!\perp \mathrm{g} \mid \mathrm{d}$, i.e., $P(\mathrm{a}, \mathrm{g}|\mathrm{d}) = P(\mathrm{a}|\mathrm{d})P(\mathrm{g}|\mathrm{d})$.

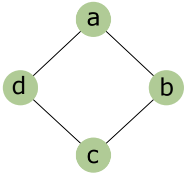
- **Proof:**

$$
\begin{aligned}
P(\mathrm{a}, \mathrm{g}|\mathrm{d}) &\propto \sum_{\mathrm{b,c,e,f}} P(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}, \mathrm{f}, \mathrm{g}) \\
&= \sum_{\mathrm{b,c,e,f}} \phi_1(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi_2(\mathrm{b}, \mathrm{c}, \mathrm{d})\phi_3(\mathrm{d}, \mathrm{e}, \mathrm{f})\phi_4(\mathrm{e}, \mathrm{f}, \mathrm{g}) \\
&= \underbrace{\sum_{\mathrm{b,c}} \phi_1(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi_2(\mathrm{b}, \mathrm{c}, \mathrm{d})}_{f(\mathrm{a,d})} \underbrace{\sum_{\mathrm{e,f}} \phi_3(\mathrm{d}, \mathrm{e}, \mathrm{f})\phi_4(\mathrm{e}, \mathrm{f}, \mathrm{g})}_{g(\mathrm{d,g})}
\end{aligned}
$$

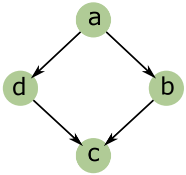
Then using the "math fact" we have $P(\mathrm{a}, \mathrm{g}|\mathrm{d}) = P(\mathrm{a}|\mathrm{d})P(\mathrm{g}|\mathrm{d})$

Probabilistic Graph Semantics

- Actually, both MN and BN have limitations



In MN "semantics",
$a \perp\!\!\!\perp c \mid b, d$ and $b \perp\!\!\!\perp d \mid a, c$,
which cannot be captured by a BN



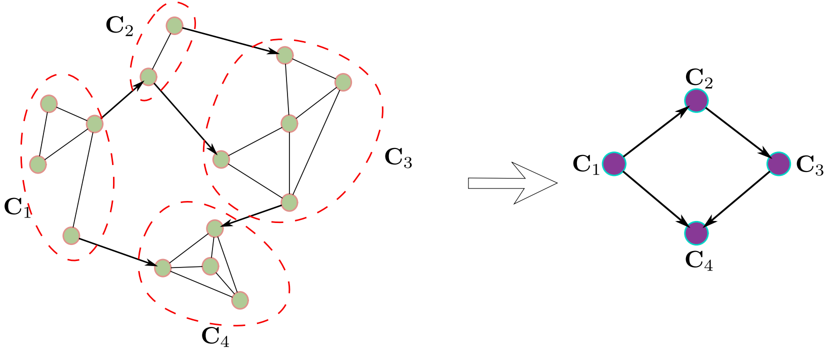
In BN "semantics",
$b \perp\!\!\!\perp d \mid a$ and $b \not\perp\!\!\!\perp d \mid a, c$,
which cannot be captured by a MN

- MNs can naturally encode "cooperative behaviour"

- BNs can naturally encode "directed behaviour"

Chain Graphical Models (BRML § 4.3)

- Chain Graphs contain both directed and undirected links, so that merging these two types of semantics
- But most engineers just treat subset of nodes $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, and $\mathcal{C}_4$ as vectors of random variables



Markov Random Field (MRF)

$$P(\mathcal{X}) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

- A MRF is a positive MN, i.e., $\mathrm{MN}^+ \triangleq \mathrm{MRF}$
- For a MN, $P(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \triangleq [\mathrm{x}_1, \ldots, \mathrm{x}_N]^\mathsf{T} \in \mathcal{X}$, i.e., $\phi_c(\mathbf{x}_c) \geqslant 0, \forall \mathbf{x}_c \in \mathcal{X}_c, c = 1, \ldots, C$
- For a MRF, $P(\mathbf{x}) > 0, \forall \mathbf{x} \triangleq [\mathrm{x}_1, \ldots, \mathrm{x}_N]^\mathsf{T} \in \mathcal{X}$, i.e., $\phi_c(\mathbf{x}_c) > 0, \forall \mathbf{x}_c \in \mathcal{X}_c, c = 1, \ldots, C$

MRF - Gibbs Distribution

$$P(\mathcal{X}) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

- For a MRF, since $\phi_c(\mathbf{x}_c) > 0, \forall \mathbf{x}_c \in \mathcal{X}_c, c = 1, \ldots, C$, we define an Energy Function (a.k.a. Potential Energy Function, Effort Function or Loss Function) on each clique by

$$E_c(\mathcal{X}_c) \triangleq -\ln \phi_c(\mathcal{X}_c)$$

  This allows us to define the "total energy" for $\mathcal{X}$ as

$$E(\mathcal{X}) \triangleq \sum_{c=1}^{C} E_c(\mathcal{X}_c)$$

- This results in the Gibbs distribution of equilibrium statistical physics via $\phi_c(\mathcal{X}_c) = e^{-E_c(\mathcal{X}_c)}$; we can also rewrite the probabilistic distribution as $P(\mathcal{X}) = \frac{1}{Z} e^{-E(\mathcal{X})}$, where $Z = \sum_{\mathcal{X}} e^{-E(\mathcal{X})}$

MRF Algorithms

- MRF yields many algorithms with various applications:
  - Simulated Annealing (SA) for stochastic optimizaton (via Markov Chain Monte Carlo (MCMC) sampling)
  - MRF image de-noising (see Bishop §8.3.3)
  - Boltzmann Machine (BM), Restricted Boltzmann Machine (RBM) and Deep RBM (D-RBM) for stochastic Neural Network (NN)
- The framework also provides a mathematical foundation for theoretical investigations into the behaviour of stochastic Deep Generative Models, such as GANs