Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 5: Undirected GMs

January 30, 2025

Reading: See course homepage

Logistics

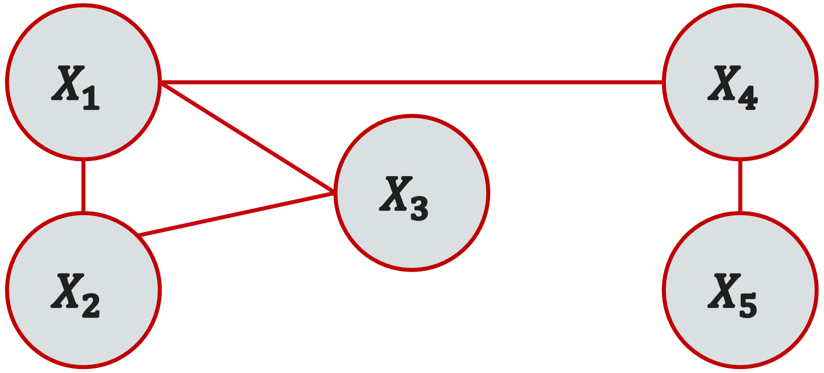
- **No class 2/11**
- HW2 deadline pushed to **2/11 11:59pm**
- **Quiz in-class on 2/13**
    - Quiz format: 3 HW problems, 2 new problems

Today

- Undirected Graphical Models
  - Markov Random Fields
  - Restricted Boltzmann Machines
  - Conditional Random Fields

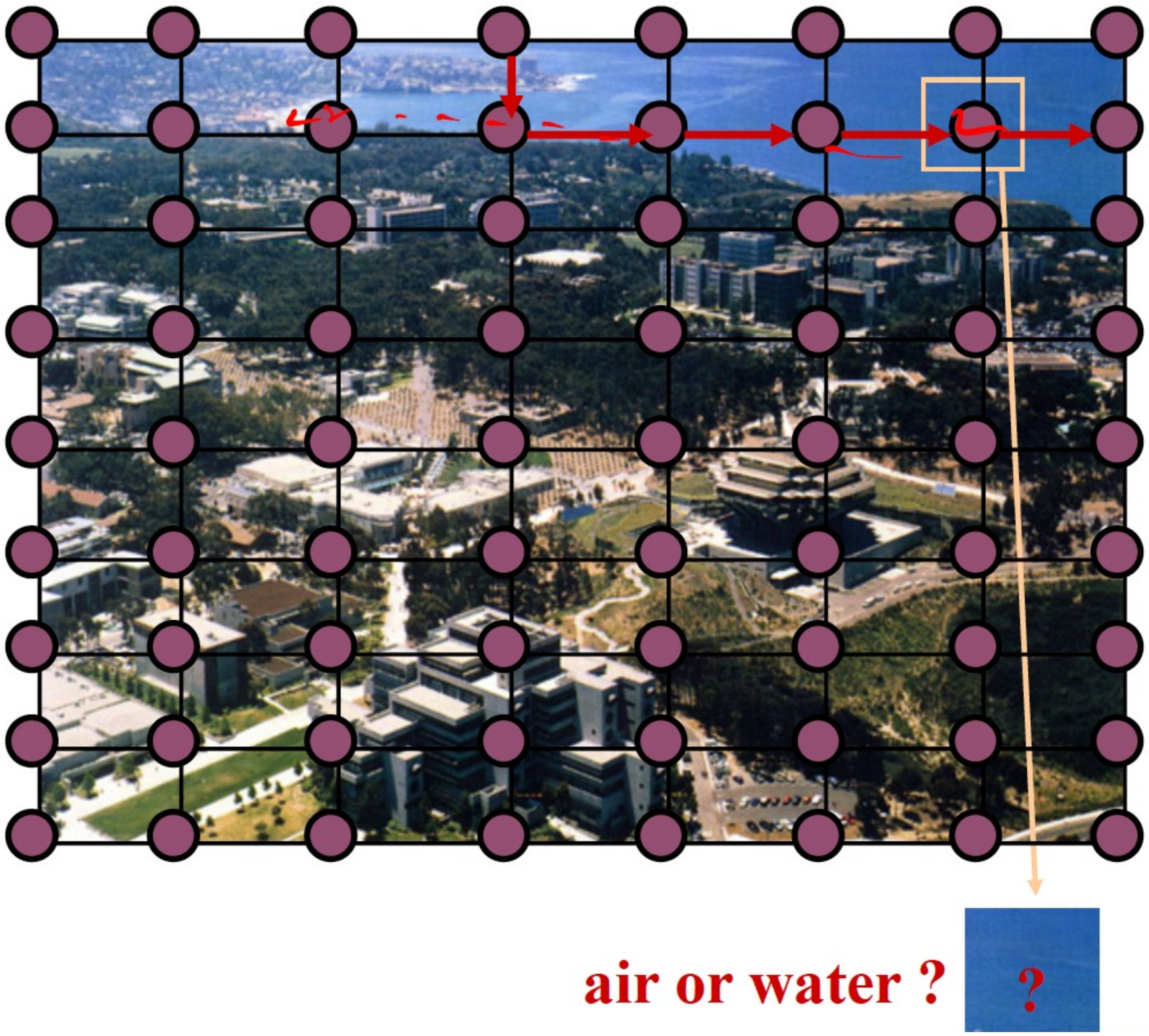
Undirected Graphical Models

Undirected Graphical Models

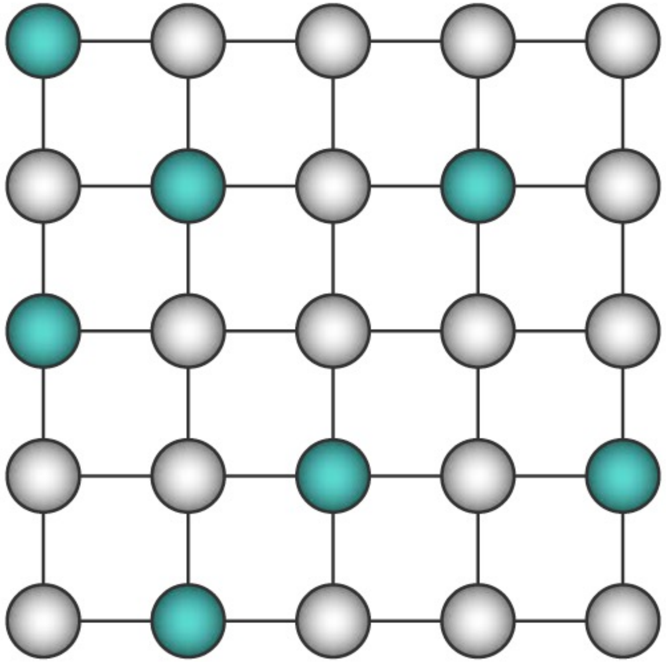


- Pairwise relationships
- No explicit way to generate samples
- Contingency constraints on node configurations

Example: Lattice



Example: Lattice



- Naturally arises in image processing, lattice physics, etc
- The states of adjacent / nearby nodes are coupled due to pattern continuity, electro-magnetic force, etc.

Representing Undirected Graphical Models

- An ***undirected graphical model*** represents a distribution $P(X)$ defined by an undirected graph H and a set of positive ***potential functions*** ψ associated with the <u>cliques</u> of H such that:

$$\mathrm{P}(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}) = \frac{1}{\mathrm{Z}} \prod_c \psi_c(X_C)$$

"Gibbs distribution"

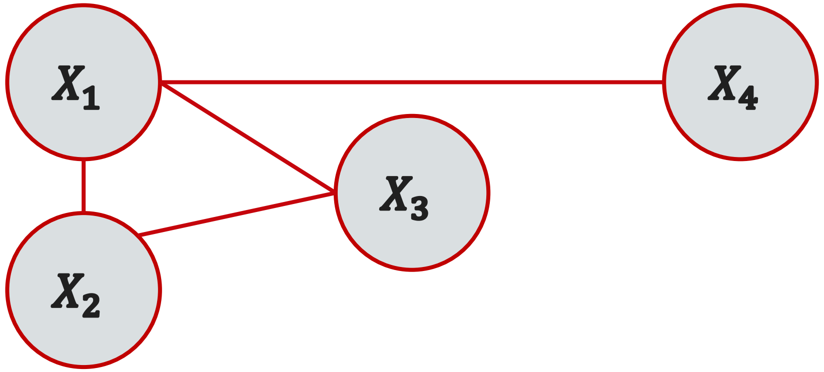
- where Z represents the **partition function**: $Z = \sum_X \prod_c \psi_c(X_c)$
- The potential function can be understood as a "score" of the joint configuration

Are $\psi_c(X_c)$ probability densities?

Is $P(X)$ a proper probability density?

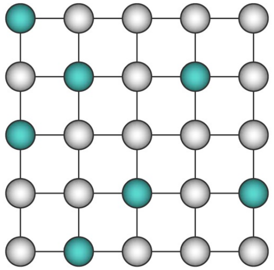
What is a clique?

- For $G = \{V, E\}$, a clique (complete subgraph) is a subgraph $G' = \{V' \subseteq V, E' \subseteq E\}$ such that nodes in V' are **fully connected**.
- A **maximal** clique is a clique such that any superset $V'' \supset V$ is *not* a clique.



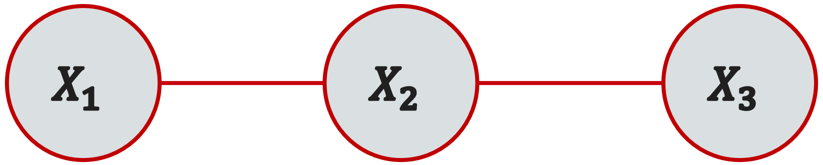
Maximal clique: $\{X_1, X_2, X_3\}$
Sub-cliques: $\{X_1, X_2\}, \{X_2, X_3\}, \{X_1, X_3\}, \{X_1, X_4\}, \{X_1\}, \{X_2\}, \{X_3\}, \{X_4\}$

Example Lattice: Ising Model from Physics



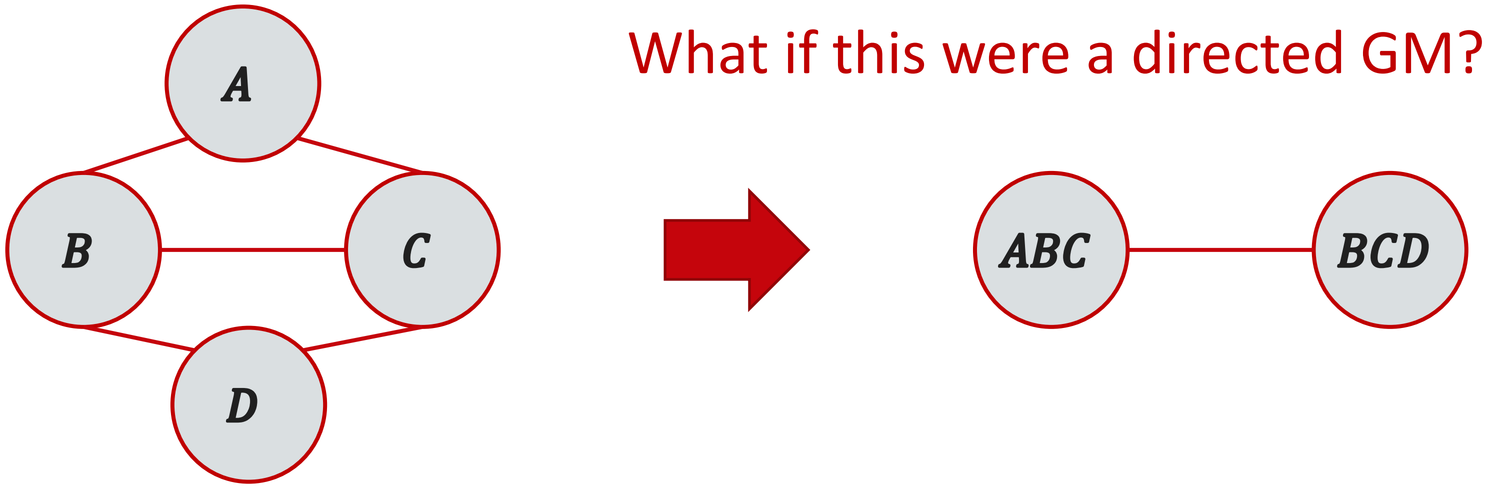
- Used to describe ferromagnetism
- Each node i has a spin variable $X_i \in \{-1, +1\}$
- Let potential function for an edge (i, j) be $\psi_{ij}(X_i, X_j) = \exp(J_{ij} X_i X_j)$ (neighboring states share spins with some strength)
- $P(X) = \frac{1}{Z} \prod_c \psi_c(X_c) = \frac{1}{Z} \exp\{\sum_{i,j} \psi_{ij}(X_i, X_j)\}$

Interpretation of Clique Potentials



- This model implies $X_1 \perp X_3 | X_2$, so joint must factorize as:
$$P(X_1, X_2, X_3) = P(X_2)P(X_1|X_2)P(X_3|X_2)$$
- We could write as $P(X_1, X_2)P(X_3 \mid X_2)$ or $P(X_2, X_3)P(X_1 \mid X_2)$, but:
  - Cannot have all potentials be marginals
  - Cannot have all potential be conditionals
- Clique potentials can be thought of as general “compatibility” of their variables, but not as probability distributions.

Example UGM: Maximal Cliques



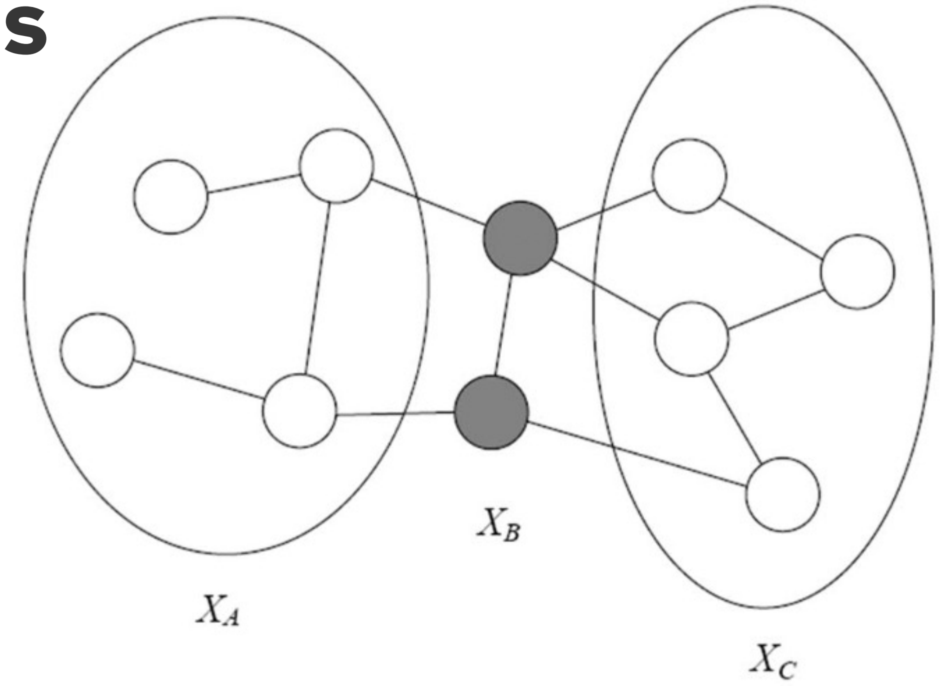
What if this were a directed GM?

$$P(A,B,C,D) = \frac{1}{Z}\psi_{ABC}(A,B,C)\psi_{BCD}(B,C,D)$$

$$Z = \sum_{ABCD} \psi_{ABC}(A,B,C)\psi_{BCD}(B,C,D)$$

Global Markov Independencies

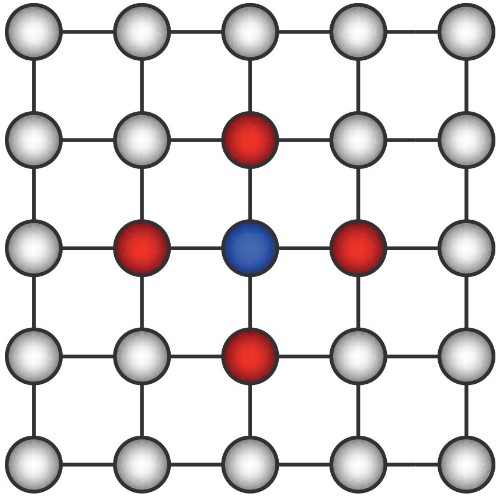
- Let H be an undirected graph:



- B separates A and C if every path from a node in A to a node in C passes through a node in B: $sep_H(A; C \mid B)$
- A probability distribution satisfies the **global Markov property** if for any disjoint A, B, C such that B separates A and C, A is independent of C given B: $I(H) = \{A \perp C \mid B: sep_H(A; C \mid B)\}$

Local Markov Independencies

- For each node X_i there is a unique Markov blanket of X_i, denoted MB_{X_i}, which is the set of neighbors of X_i in the graph.



- The ***local Markov independencies*** in H are:

$$I_l(H) = \{X_i \perp V - \{X_i\} - MB_{X_i} \mid MB_{X_i} : \forall\, i\}$$

- In other words, X_i is independent of the rest of the nodes in the graph given its immediate neighbors.

Pairwise Markov Independencies

- The pairwise Markov independencies associated with H are:

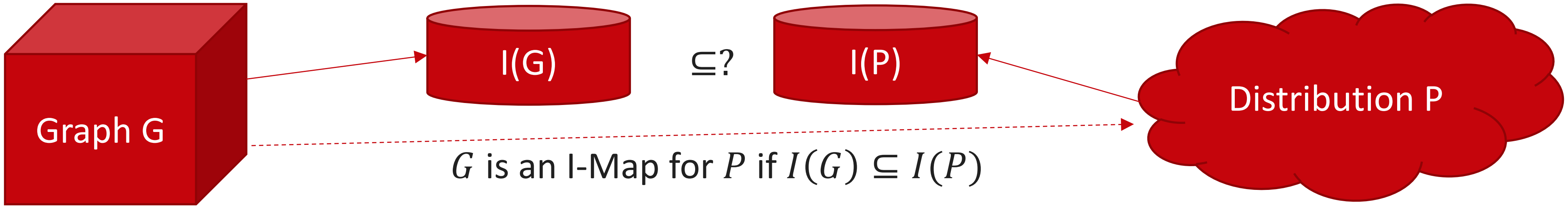
$$I_P(H) = \{ X \perp Y \mid V \setminus \{X, Y\} : \{X, Y\} \notin E \}$$

e.g.

$X_1 \perp X_5 | \{X_2, X_3, X_4\}$

Recall: I-Maps

- Independence set: Let P be a distribution over X. We define $I(P)$ to be the set of independences $(X \perp Y \mid Z)$ that hold in P.
- I-Map: Let G be any graph object with an associated independence set $I(G)$. We say that G is an **I-map** for an independence set I if $I(G) \subseteq I$.
- I-Map of Distribution: We say G is an I-map for P if G is an I-map for $I(P)$, when we use $I(G)$ as the associated independence set.

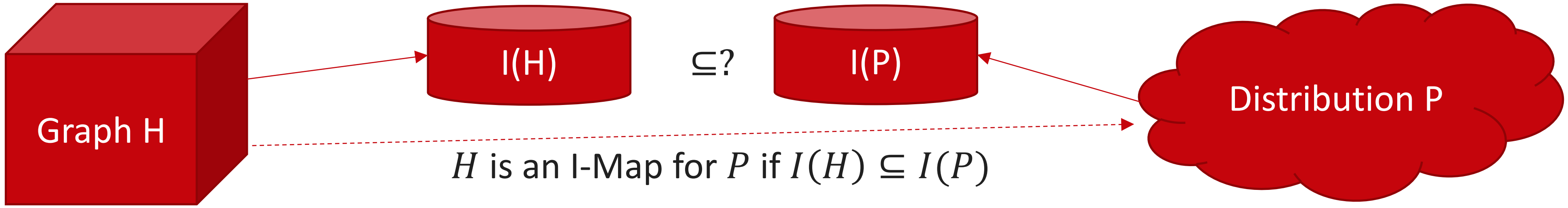


I-Maps of UG

- An UG H is an I-Map for a distribution P if $I(H) \subseteq I(P)$
- P is a ***Gibbs Distribution*** over H if it can be represented as:

$$P(X) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c)$$

- Theorem (soundness): If P is a Gibbs Distribution over H, then H is an I-Map of P.

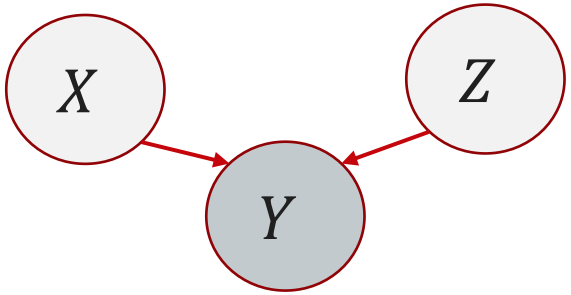


Perfect Maps

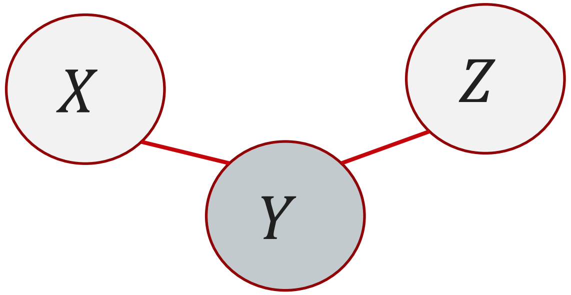
- An UG H is a ***perfect map*** for P if for any X, Y, Z, we have that

$$sep_H(X; Z \mid Y) \leftrightarrow X \perp Z \mid Y$$

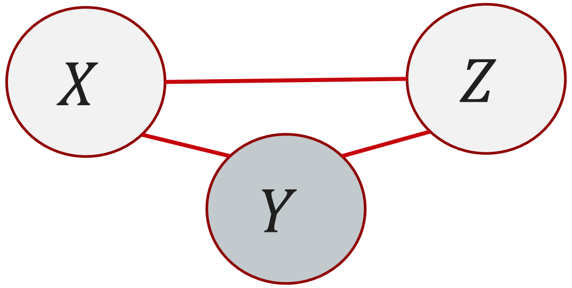
- Not every distribution has a perfect map as an UG.
  - Example: V-structure



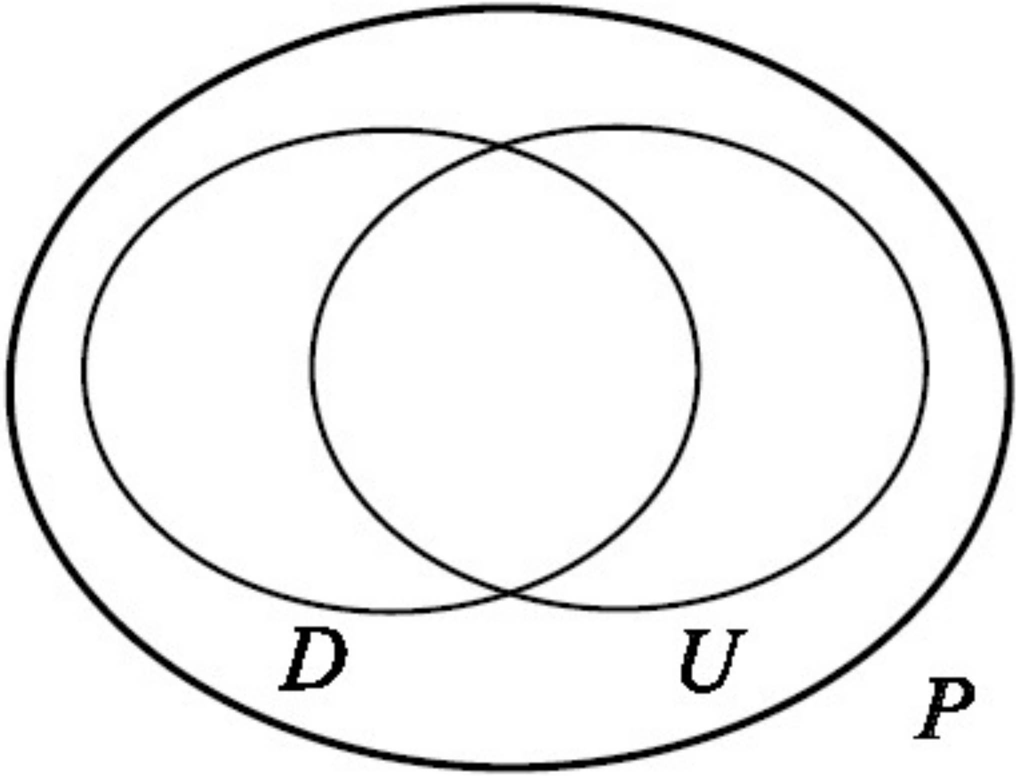
$$X \perp Z$$
$$\neg X \perp Z \mid Y$$



$$X \perp Z$$
$$X \perp Z \mid Y$$



GMs and UGMs rep. overlapping sets of dists



Exponential Families

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive).
- We can represent a clique potential ψ in an **unconstrainted** form using a real-valued "energy" function φ and have:

$$\psi_c(X_c) = \exp\big(-\phi_c(X_c)\big)$$

- This gives the joint a nice additive structure:

$$P(X) = \frac{1}{Z}\exp\left\{-\sum_{c\in C}\phi_c(X_c)\right\} = \frac{1}{Z}\exp\{-\underbrace{H(X)}_{\text{"Energy"}}\}$$

In physics, this is called the **Boltzmann distribution**.
In statistics, this is called a **log-linear model**.

Aside: MAP Inference = Free Energy Minimization

$$P_{Boltzmann}(X) = \text{argmin}_H F\big(P(X; H)\big) = \text{argmin}_H \mathrm{E}[H(X)] - TS\big(P(X)\big)$$

Distribution observed in nature

Free energy

Expected energy

Entropy at temp T

Negative log likelihood

Entropy = Uncertainty - LogPrior

$$Q^*(\theta) = \text{argmin}_Q F(Q) = \text{argmin}_Q \mathrm{E}_{Q(\theta)}[-\log P(X \mid \theta)] - TS(Q) - \mathrm{E}_{Q(\theta)}[\log P(\theta)]$$

Will make more sense after we study variational inference

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON