Lecture 4: Undirected Graphical Models

Department of Biostatistics
University of Michigan
zhenkewu@umich.edu
http://zhenkewu.com/teaching/graphical_model

15 September, 2016

Lecture 3 Main Points Again

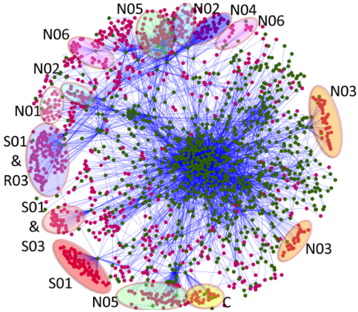
Representation of Directed Acyclic Graphs (DAG)

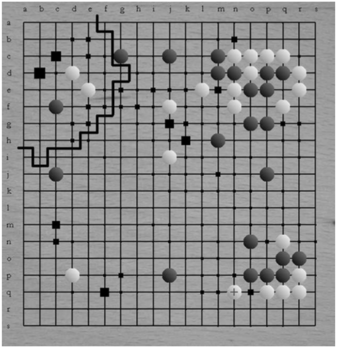
- *Motivation*: Need a system that can
  - Clearly represent human knowledge about informational relevance
  - Afford qualitative and robust reasoning
- *Representation*:
  - Connect d-separation (graphical concept) to conditional independence (probability concept)
  - Directed edges (arrows) encode *local* dependencies
- Not every joint probability distribution has a DAG with exactly the same set of conditional independencies (represented by the d-separation triplets from the DAG).
- Reading (optional): Pearl and Verma (1987). *The logic of representing dependencies by directed acyclic graphs.*

Undirected Graphical Models

- DAGs using directed edges to guide the specification of components in the joint probability distributions: $[X_1, \ldots, X_p] = \prod_j [X_j \mid Pa^{\mathcal{G}}_{X_j}]$ (local Markov condition)
- **Undirected graphical (UG)** models also provide another system for qualitatively representing vertex-dependencies, esp. when the directionality of interactions are unclear; Gives correlations
- Also known as: Markov Random Field (MRF), or Markov network
- Rich applications in spatial statistics (spatial interactions), natural language processing (word dependencies), network discoveries (e.g., neuron activation patterns, protein interaction networks),...

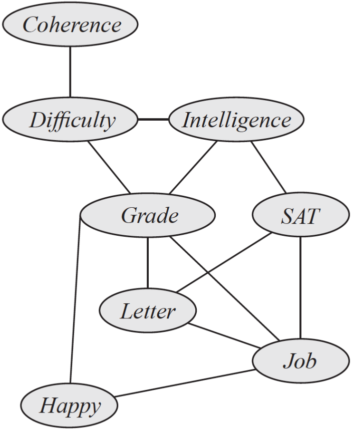
UG Examples (Protein Networks and Game of Go)





Stern et al. (2004), Proceedings of 23rd ICML

Undirected Graphical Models



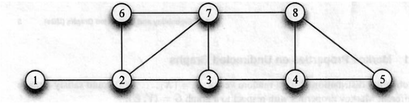
- Pairwise non-causal relationships
- Can readily write down the model, but not obvious how to generate samples from it

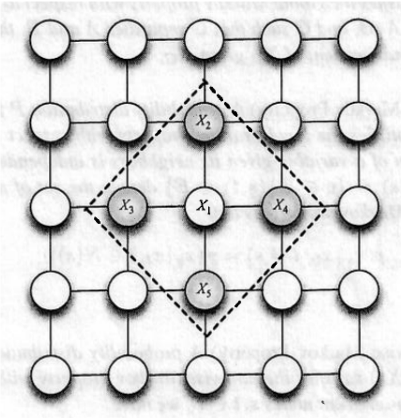
Markov Properties on UG

A probability distribution P for a random vector $X = (X_1, \ldots, X_d)$ could satisfy a range of different Markov properties with respect to a graph $G = (V, E)$, where V is the set of vertices, each corresponding to one of $\{X_1, \ldots, X_d\}$, and E is the set of edges.

- **Global Markov Property:** P satisfies the global Markov property with respect to a graph G if for any disjoint vertex subsets A, B, and C, such that C *separates* A and B, the random variables X_A are conditionally independent of X_B given X_C.
- Here, we say C separates A and B if every path from a node in A to a node in B passes through a node in C.
- **Local Markov Property:** P satisfies the local Markov property with respect to G if the conditional distribution of a variable given its neighbors is independent of the remaining nodes.
- **Pairwise Markov Property:** P satisfies the pairwise markov property with respect to G if for any pair of non-adjacent nodes, $s, t \in V$, we have $X_s \perp X_t \mid X_{V \setminus \{s,t\}}$

Separation





Relationships of Different Markov Properties

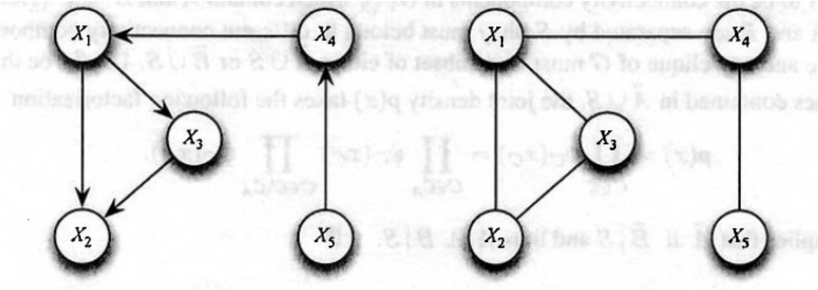
A distribution that satisfies the global Markov property is said to be a Markov random field or Markov network with respect to the graph.

- **Proposition 1**: For any undirected graph G and any distribution P, we have:

$$\boxed{\text{global Markov property} \implies \text{local Markov Property} \implies \text{pairwise Markov property}}$$

- **Proposition 2**: If the joint density $p(\boldsymbol{x})$ of the distribution P is positive and continuous with respect to a product measure, then pairwise Markov property implies global Markov property.

Therefore, for distributions with positive continuous densities, the global, local, and pairwise Markov properties are **equivalent**.
We usually say a distribution P is **Markov** to G, if P satisfies the global Markov property with respect to a graph G.



- Unlike a DAG that encodes factorization by conditional probability distributions, UG does this in terms of **clique potentials**, where *clique* in a graph is a fully connected subset of vertices.
- A clique is a *maximal clique* if it is not contained in any larger clique.

Factorization by Clique Decompositions

- Let $\mathcal{C}$ be a set of all maximal cliques in a graph. A probability distribution factorizes with respect to this graph G if it can be written as a product of factors, one for each of the maximal cliques in the graph:
$$p(x_1, \ldots, x_d) = \prod_{c \in \mathcal{C}} \psi_C(x_C).$$
- Similarly, a set of clique potentials $\{\psi_C(x_C) \geq 0\}_{C \in \mathcal{C}}$ determines a probability distribution that factors with respect to the graph G by normalizing:
$$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_C(x_C).$$
- The normalizing constant, or partition function **Z** sums or integrates over all settings of the random variables. Note that **Z** may contain parameters from the potential functions.

Factorization and Markov Property

- **Theorem 1:** For any undirected graph $G = (V, E)$, a distribution P that factors with respect to the graph also satisfies the global Markov property on the graph.
- **Next question: under what conditions the Markov properties imply factorization with respect to a graph?**
- **Theorem (Hammersley-Clifford-Besag; Discrete Version)**. Suppose that $G = (V, E)$ is a graph and $X_i, i \in V$ are random variables that take on a finite number of values. If $\mathbb{P}(x) > 0$ is strictly positive and satisfies the local Markov property with respect to G, then it factorizes with respect to G.

Factorization and Markov Property (continued)

- For positive distributions,
  Global Markov $\Leftrightarrow$ Local Markov $\Leftrightarrow$ Factorization

Comment

- Next lecture: learn the relationships between DAGs and UGs; when can we convert a DAG to an UG; how can we do it? (Hint: moralization; important for posterior inference)
- **Reading**: Section 4.5, Koller and Friedman (2009)