ECE 175B: Probabilistic Reasoning and Graphical Models
Lecture 4: Bayesian Belief Networks

Florian Meyer & Ken Kreutz-Delgado

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Factorizing Joint Distributions

- Complexity of Joint Probability Distributions: Let $\mathcal{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$ be a set of N random variables with joint pmf $P(\mathcal{X}) = P(\mathrm{x}_1, \ldots, \mathrm{x}_N)$
  - assume each random variable x_i is categorically taking $k_i \geqslant 2$ values; how many possible values does $\mathcal{X}$ have?
  - imagine $k_i = 2$, and $N = 300$, then at least $2^{300} - 1 \approx 10^{90}$ values (v.s. number of atoms in the universe $10^{78} \sim 10^{82}$)
- To address this issue we exploit the structure of $P(\mathcal{X})$
  - factorize $P(\mathcal{X})$ based on the product rule (a.k.a. chain rule) and make use of the conditional independencies that exists among the random variables $\mathrm{x}_1, \ldots, \mathrm{x}_N$
  - encode the resulting factorization of $P(\mathcal{X})$ by a directed acyclic graph (DAG) called a Bayes Network or a Belief Network (BN)

The Chain Rule

- We can factorize the joint probability distribution $P(\mathcal{X})$ as product of conditional distributions

$$\begin{aligned} P(\mathcal{X}) &= P(\mathrm{x}_1, \dots, \mathrm{x}_N) \\ &= P(\mathrm{x}_N | \mathrm{x}_{N-1}, \dots, \mathrm{x}_1) \cdots P(\mathrm{x}_j | \mathrm{x}_{j-1}, \dots, \mathrm{x}_1) \cdots P(\mathrm{x}_1) \\ &= \prod_{j=1}^{N} P(\mathrm{x}_j | \mathrm{x}_{j-1}, \dots, \mathrm{x}_1) \end{aligned}$$

- Note that there are $N!$ different ways to factorizing the joint distribution (each corresponding to a permutation of $\mathrm{x}_1, \dots, \mathrm{x}_N$)
- To simplify the notation we just relabel any permutation $\mathrm{x}_{i_1}, \dots, \mathrm{x}_{i_N}$ as $\mathrm{x}_1 \leftarrow \mathrm{x}_{i_1}, \dots, \mathrm{x}_N \leftarrow \mathrm{x}_{i_N}$ such that we have ancestral ordering, i.e., the last distribution in the chain is always $P(\mathrm{x}_1)$

Markov Property of Conditional Probability Distributions

- It is still equally challenging to represent the conditional probability distribution $P(\mathrm{x}_j|\mathrm{x}_{j-1},\ldots,\mathrm{x}_1)$ for a large j
- Question: Do we really need to include the whole prior set $\mathbf{pr}(\mathrm{x}_j) = \{\mathrm{x}_{j-1},\ldots,\mathrm{x}_1\}$?
- Answer: If there is conditional statistical independence, we only need to consider the smaller parents set of x_j, denoted as $\mathbf{pa}(\mathrm{x}_j)$

Markov Property of Conditional Probability Distributions

- Define the parents of x_j, denoted $\mathbf{pa}(\mathrm{x}_j)$, to be the smallest subset of $\mathbf{pr}(\mathrm{x}_j)$, i.e.,

$$\mathbf{pa}(\mathrm{x}_j) \subset \mathbf{pr}(\mathrm{x}_j) \subset \mathcal{X}$$

  such that

$$P\big(\mathrm{x}_j | \mathbf{pa}(\mathrm{x}_j)\big) = P\big(\mathrm{x}_j | \mathbf{pr}(\mathrm{x}_j)\big)$$

- This implies that

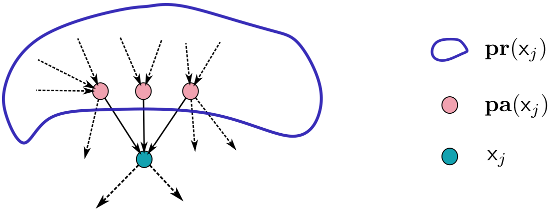
$$\mathrm{x}_j \perp\!\!\!\perp \big(\mathbf{pr}(\mathrm{x}_j) \backslash \mathbf{pa}(\mathrm{x}_j)\big) \bigm| \mathbf{pa}(\mathrm{x}_j)$$

- Now we can write $P(\mathcal{X}) = \prod_{j=1}^{N} P(\mathrm{x}_j | \mathbf{pa}(\mathrm{x}_j))$ and encode this factorization by a DAG

Build a BN from the factorization of $P(\mathcal{X}) = P\big(\mathrm{x}_j|\mathbf{pa}(\mathrm{x}_j)\big)$

Algorithm 1: Build a BN from the factorization of $P(\mathcal{X})$

$[\mathcal{G} = (\mathcal{V}, \mathcal{E})] = \mathrm{BN}\big[\prod_{j=1}^{N} P\big(\mathrm{x}_j|\mathbf{pa}(\mathrm{x}_j)\big)\big]$
Initialize $\mathcal{V} = \emptyset$ and $\mathcal{E} = \emptyset$;
for $j = 1 : N$ **do**
    $j \leftarrow \mathrm{x}_j$ `// introduce a new node`
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{j\}$ `// add the new node to the vertex set`
    for $\mathrm{x}_i \in \mathbf{pa}(\mathrm{x}_j)$ **do**
        $(i, j) \leftarrow (\mathrm{x}_i, \mathrm{x}_j)$ `// draw edge from parent to the node`
        $\mathcal{E} \leftarrow \mathcal{E} \cup \{(i, j)\}$ `// add new edge to the edge set`
    end
end



Ancestral Order of a BN

- Once a BN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed corresponding to the factorization $P(\mathcal{X}) = \prod_{j=1}^{N} P\big(\mathrm{x}_j | \mathbf{pa}(\mathrm{x}_j)\big)$, we can define $\mathcal{X}$-subsets of interest in terms of the corresponding subsets of $\mathcal{V}$

  - $h \in \mathbf{anc}(j) \iff \mathrm{x}_h \in \mathbf{anc}(\mathrm{x}_j) \qquad (h < j)$
  - $i \in \mathbf{pa}(j) \iff \mathrm{x}_i \in \mathbf{pa}(\mathrm{x}_j) \qquad (i < j)$
  - $k \in \mathbf{child}(j) \iff \mathrm{x}_k \in \mathbf{child}(\mathrm{x}_j) \qquad (k > j)$
  - $l \in \mathbf{desc}(j) \iff \mathrm{x}_l \in \mathbf{desc}(\mathrm{x}_j) \qquad (l > j)$

- Note that for $\mathrm{x}_h, \mathrm{x}_i, \mathrm{x}_k, \mathrm{x}_l$ related as defined above, we have $h \leqslant i < k \leqslant l$, i.e., we have ancestral ordering of the nodes of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Note that $\mathbf{pa}(\mathrm{x}_j) \subseteq \mathbf{anc}(\mathrm{x}_j) \subseteq \mathbf{pr}(\mathrm{x}_j) \subset \mathcal{X}$ (why?) and thus $\mathrm{x}_j \perp\!\!\!\perp \big(\mathbf{anc}(\mathrm{x}_j) - \mathbf{pa}(\mathrm{x}_j)\big) \mid \mathbf{pa}(\mathrm{x}_j)$

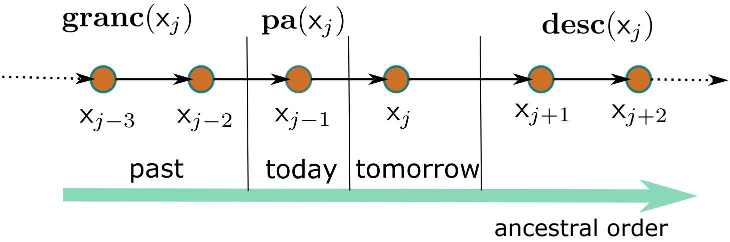
Ancestral Order of a BN (cont.)

- To make this clearer, define the "grand-ancestors" of x_j as $\mathbf{granc}(\mathrm{x}_j) \triangleq \mathbf{anc}(\mathrm{x}_j) \backslash \mathbf{pa}(\mathrm{x}_j)$ with $\mathbf{anc}(\mathrm{x}_j) = \mathbf{granc}(\mathrm{x}_j) \cup \mathbf{pa}(\mathrm{x}_j)$

- Then we get the Basic Markov Property as

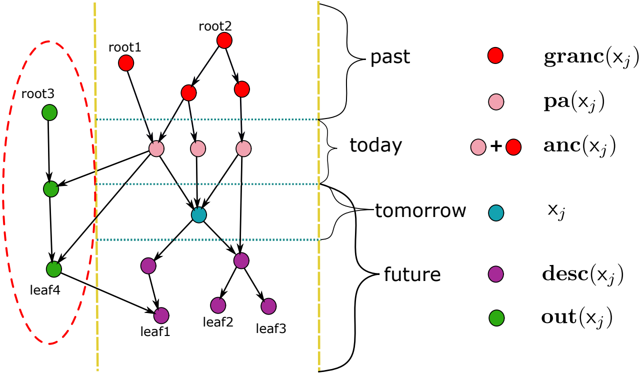
$$P(\mathrm{x}_j|\mathbf{pa}(\mathrm{x}_j), \mathbf{granc}(\mathrm{x}_j)) \Longleftrightarrow \mathrm{x}_j \perp\!\!\!\perp \mathbf{granc}(\mathrm{x}_j) \mid \mathbf{pa}(\mathrm{x}_j)$$

- Example: In a serial chain where the index $i \in \mathcal{V}$ represents time, "tomorrow" is independent of "past" given "today"



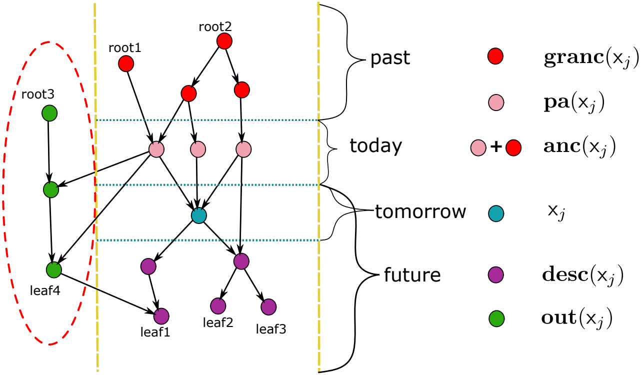
Ancestral Order of a BN (cont.)

- Note that the ancestor/descedent relationship related to x_j restricts a timeline w.r.t. x_j
  - all ancestors of x_j have at least one directed path from it to x_j
  - all descedents of x_j have at least one directed path from x_j to itself
- For nodes that do not share a directed path with x_j, we say they are "out of the timeline" of x_j



Ancestral Order of a BN (cont.)

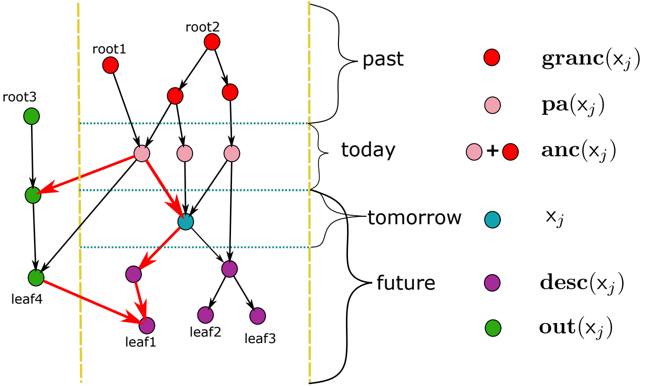
- All nodes that are out of the timeline of x_j are denoted as $\mathbf{out}(x_j)$
- The ordering of $x \in \mathbf{out}(x_j)$ w.r.t x_j is unclear



- In a directed tree, every node is "in the timeline" of every other node

Ancestral Order of a BN (cont.)

- Note that in general for $x \in \mathbf{out}(x_j)$, we have $\mathbf{pa}(x) \cap \mathbf{pa}(x_j) \neq \emptyset$
- Non-empty intersections also exists for **anc**, **child**, and **desc**



Finding the Best BN

- Recall that for a certain joint distribution of a set of N random variables
  - there are $N!$ rearrangements of the factorizing order
  - we can build a BN corresponding to a specific factorization
- Questions: Do all factorizations represent the same causal relationship between the random variables? Do all factorizations yield the same specification complexity of conditional probability distributions? If not, which factorization should we choose?