COMP538: Introduction to Bayesian Networks
Lecture 3: Probabilistic Independence and Graph Separation

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008
Objective

- **Objective:**
  - Discusses the relationship between *probabilistic independence* and *graph separation* in Bayesian networks.
  - Given a BN structure, a DAG, what independence relationships are represented?
  - Given a joint distribution, under what conditions can the independence relationships it entails be represented using a DAG? How much?

- **Reading:** Zhang & Guo: Chapter 3
Outline

1 An Intuitive Account
   • Special Cases
   • The General Case

2 D-Separation and Independence
   • Some Lemmas
   • Proof of Main Result
   • Corollaries

3 Representing Independence using DAG
Intuitive Meaning of Independence

- Given: A Bayesian network and two variables $X$ and $Y$.
- Question:
  - Are $X$ and $Y$ independent?
  - What are the (graph-theoretic) conditions under which $X$ and $Y$ are independent?
- We will try to answer this question based on intuition.
- This exercise will lead to the concept of d-separation.

Intuitive meaning of independence:
- $X$ and $Y$ are dependent under some condition $C$ iff knowledge about one influences belief about the other under $C$.
- $X$ and $Y$ are independent under some condition $C$ iff knowledge about one does not influence belief about the other under $C$. 
Case 1: Direction connection

- If $X$ and $Y$ are connected by an edge, then $X$ and $Y$ are dependent (under the empty condition).
- Information can be transmitted over one edge.

Example:

- Burglary and Alarm are dependent:
  - My knowing that a burglary has taken place increases my belief that the alarm went off.
  - My knowing that the alarm went off increases my belief that there has been a burglary.
Case 2: Serial connection

- If $Z$ is not observed, $X$ and $Y$ are dependent.

- Information can be transmitted between $X$ and $Y$ through $Z$ if $Z$ is not observed.

- If $Z$ is observed, $X$ and $Y$ are independent.

- Information cannot be transmitted between $X$ and $Y$ through $Z$ if $Z$ is observed. Observing $Z$ blocks the information path.
Case 2: Serial connection/Example

- If A not observed, B and M call are dependent:
  - My knowing that a burglary has taken place increases my belief on Marry call.
  - My knowing that Marry called increases my belief on burglary.

- If A is observed, B and M are conditionally independent:
  - If I already know that the alarm went off,
    - My further knowing that a burglary has taken place would not increases my belief on Marry call.
    - My further knowing that Marry called would not increases my belief on burglary.
Case 3: Diverging connection (common cause)

- If $Z$ is not observed, $X$ and $Y$ are dependent.

- **Information can be transmitted through $Z$ among children of $Z$ if $Z$ is not observed.**

- If $Z$ is observed, $X$ and $Y$ are independent.

- **Information cannot be transmitted through $Z$ among children of $Z$ if $Z$ is observed. Observing $Z$ blocks the information path.**
Case 3: Diverging connection/Example

- If $A$ is not observed, $J$ and $M$ are dependent:
  - My knowing that John called increases my belief on Marry call.
  - My knowing that Marry called increases my belief on John call.

- If $A$ is observed, $J$ and $M$ are conditionally independent:
  - If I already know that the alarm went off,
    - My further knowing that John called would not increase my belief on Marry call.
    - My further knowing that Marry called would not increase my belief on John call.
Case 4: Converging connection (common effect)

- If neither $Z$ nor any of its descendant are observed, $X$ and $Y$ are independent.

- Information cannot be transmitted through $Z$ among parents of $Z$. It leaks down $Z$ and its descendants.

- If $Z$ or any of its descendant is observed, $X$ and $Y$ are dependent.

- Information can be transmitted through $Z$ among parents of $Z$ if $Z$ or any of its descendants are observed. Observing $Z$ or its descendants opens the information path.
Case 4: Converging connection/Example

- **E and B are conditionally dependent if A is observed:**
  - If I already know that the alarm went off,
    - My further knowing that there has been an earthquake decreases my belief on Burglary.
  - My further knowing that there has been a burglary decreases my belief on earthquake.
  
  **Explaining away.**

- **E and B are conditionally dependent if M is observed:**
  - Observing Marry call gives us some information about Alarm. So we are back to the previous case.

- **E and B are marginally independent (if A, M and J not observed).**
Hard Evidence and Soft Evidence

- **Hard evidence** on a variable: The value of the variable is directly observed.
- **Soft evidence** on a variable: The value of the variable is NOT directly observed. However the value of a descendant is observed.

The rules restated:

- Hard evidence blocks information path in the case of serial and diverging connection
- Both hard and soft evidence are enough for opening of information path in the case of converging connection.
A path between $X$ and $Y$ is **blocked** by a set $Z$ of nodes if

1. Either that path contains a node $Z$ that is in $Z$ and the connection at $Z$ is either serial or diverging.

2. Or that the path contains a node $W$ such that $W$ and its descendants are not in $Z$ and the connection at $W$ is a converging connection.
Suppose all variables in $Z$ are the observed variables.

Then a path between $X$ and $Y$ being blocked by $Z$ implies:

1. Either information cannot be transmitted through $Z$ because observing $Z$ blocks that path.
2. Or information cannot be transmitted through $W$, it leaks through $W$.

In both cases, information cannot be transmitted between $X$ and $Y$ along the path.

If path is not blocked, on the other hand, information CAN flow between $X$ and $Y$. 
D-separation

- Two nodes $X$ and $Y$ are **d-separated** by a set $Z$ if
  - All paths between $X$ and $Y$ are blocked by $Z$.

- Theorem 3.1:
  - If $X$ and $Y$ are d-separated by $Z$, then $X \perp Y | Z$.

- It should be pointed out that this conclusion is derived from intuition.

- One of the main tasks in this lecture is to rigorously show that the conclusion is indeed true.
Examples

- A d-separated (by empty set) from C, F, G, J
- A d-separated by \{M, B\} from G
- A d-separated by \{E, K, L\} from M
- I d-separated by \{E, K, L\} from M

Exercise: Try more examples on your own.
D-Separation and Independence

Outline

1. An Intuitive Account
   - Special Cases
   - The General Case

2. D-Separation and Independence
   - Some Lemmas
   - Proof of Main Result
   - Corollaries

3. Representing Independence using DAG
Ancestral Sets

- Let $X$ be a set of nodes in a Bayesian network.
- The **ancestral set** $an(X)$ of $X$ consists of
  - All nodes in $X$ and all the ancestors of nodes in $X$.
- Example: The ancestral set of $\{I, G\}$ consists of
  $$\{I, G, A, B, C, D, E\}$$
- We say that $X$ is **ancestral** if
  $$X = an(X)$$
- A **leaf node** is one without children. Examples: $M, L$
A Lemma

Lemma (3.1)

Suppose $\mathcal{N}$ is a Bayesian network, and $Y$ is a leaf node. Let $\mathcal{N}'$ be the Bayesian network obtained from $\mathcal{N}$ by removing $Y$. Let $X$ be the set of all nodes in $\mathcal{N}'$. Then

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}'}(X).$$

Proof

$$P_{\mathcal{N}}(X) = \sum_Y P_{\mathcal{N}}(X, Y)$$

$$= \sum_Y \left[ \prod_{W \in X} P(W|\text{pa}(W)) \right] P(Y|\text{pa}(Y))$$

$$= \prod_{W \in X} P(W|\text{pa}(W)) \sum_Y P(Y|\text{pa}(Y))$$

$$= \prod_{W \in X} P(W|\text{pa}(W))$$

$$= P_{\mathcal{N}'}(X)$$
A Lemma

- The third equality is true because, being a leaf node, \( Y \) is not in \( X \) and cannot be in any \( pa(W) \) for any \( W \in X \).

- The fourth equality is true because probability sum to one. Q.E.D
First Proposition

**Proposition (3.1)**

Let $\mathbf{X}$ be a set of nodes in a Bayesian network $\mathcal{N}$. Suppose $\mathbf{X}$ is ancestral. Let $\mathcal{N}'$ be the Bayesian network obtained from $\mathcal{N}'$ by removing all nodes outside $\mathbf{X}$. Then,

$$P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X}).$$

**Proof:**

- Consider the following procedure
  - While there are nodes outside $\mathbf{X}$,
    - Find a leaf node. (There must be one. Exercise.)
    - Remove it.
  - Afterwards, we get $\mathcal{N}'$.
- And according to Lemma 3.1, the probability distribution of $\mathbf{X}$ remains unchanged throughout the procedure.
- The proposition is hence proved. Q.E.D.
Second Proposition

Proposition (3.2)

Let \(X, Y,\) and \(Z\) be three disjoint sets of nodes in a Bayesian network such that their union is the set of all nodes.

- If \(Z\) d-separates \(X\) and \(Y,\) then

\[
X \perp Y | Z
\]

Proof:

- Let \(Z_1\) be the set of nodes in \(Z\) that have parents in \(X.\) And let \(Z_2 = Z \setminus Z_1.\)

- Because \(Z\) d-separates \(X\) and \(Y,\)
  - For any \(W \in X \cup Z_1,\)
    \[pa(W) \subseteq X \cup Z.\]
  - For any \(W \in Y \cup Z_2,\)
    \[pa(W) \subseteq Y \cup Z.\]
Proof Second Proposition (cont’d)

Consider

\[
P(X, Z, Y) = \prod_{W \in X \cup Z \cup Y} P(W|\text{pa}(W))
\]

\[
= \left[ \prod_{W \in X \cup Z_1} P(W|\text{pa}(W)) \right] \left[ \prod_{W \in Z_2 \cup Y} P(W|\text{pa}(W)) \right]
\]

Note that

- \( \prod_{W \in X \cup Z_1} P(W|\text{pa}(W)) \) is a function of \( X \) and \( Z \)
- \( \prod_{W \in Z_2 \cup Y} P(W|\text{pa}(W)) \) is a function of \( Z \) and \( Y \).

It follows from Proposition 1.1 (of Lecture 1) that

\[
X \perp Y|Z
\]

Q.E.D
Global Markov property

Theorem (3.1)

Given a Bayesian network, let $X$ and $Y$ be two variables and $Z$ be a set of variables that does not contain $X$ or $Y$. If $Z$ d-separates $X$ and $Y$, then

$$X \perp Y | Z$$

Proof:

- Because of Proposition 3.1, we can assume that $\text{an} \{X, Y \} \cup Z$ equals the set of all nodes.
  - $X \perp Y | Z$ in original network iff it is true in the restriction onto the ancestral set.
  - $Z$ d-separates $X$ and $Y$ in original network iff it is true in the restriction onto the ancestral set. (Exercise)
Proof of Global Markov property (cont’d)

- Let $X$ be the set of all nodes that are NOT d-separated from $X$ by $Z$.
- Let $Y$ be the set of all nodes that are neither in $X$ or $Z$.
- Because of Proposition 3.2, $X \perp Y | Z$.
- Because of Proposition 1.1, there must exist functions $f(X, Z)$ and $g(Z, Y)$ such that
  \[ P(X, Z, Y) = f(X, Z)g(Z, Y) \]
- Note that $X \in X$ and $Y \in Y$.
- Let $X' = X \setminus \{X\}$ and $Y' = Y \setminus \{Y\}$.
- We have
  \[ P(X, X', Z, Y, Y') = f(X, X', Z)g(Z, Y, Y') \]
Proof of Global Markov property (cont’d)

■ Consequently

\[
P(X, Y, Z) = \sum_{X', Y'} P(X, X', Z, Y, Y')
\]

\[
= \sum_{X', Y'} f(X, X', Z)g(Z, Y, Y')
\]

\[
= [\sum_{X'} f(X, X', Z)][\sum_{Y'} g(Z, Y, Y')]
\]

\[
= f'(X, Z)g'(Z, Y)
\]

That is

\[
X \perp Y \mid Z
\]

Q.E.D
Markov blanket

- In a Bayesian network, the **Markov blanket** of a node $X$ is the set consisting of
  - Parents of $X$
  - Children of $X$
  - Parents of children of $X$

- Example:

The Markov blanket of $I$ is $\{E, H, J, K, L\}$
Corollary (3.1)

In a Bayesian network, a variable $X$ is conditionally independent of all other variables given its Markov blanket. (This is why it is so called.)

**Proof:**

- Because of Theorem 3.1, it suffices to show that
  - The Markov blanket of $X$ d-separates $X$ from all other nodes.

- This is true because, in any path from $X$ to outside its Markov blanket, the connection at that last node before leaving the blanket is either serial or diverging. Q.E.D
Local Markov property

Corollary (3.2)

(Local Markov property) In a Bayesian network, a variable $X$ is independent of all its non-descendants given its parents.

Proof:

- Because of Theorem 3.1, it suffices to show that
  - $pa(X)$ d-separates $X$ from the non-descendants of $X$.
- Consider a path between $X$ and a non-descendant $Y$. Let $Z$ be the neighbor of $X$ on the path.
  - Case 1: $Z \in pa(X)$,
    - The connection at $Z$ is not converging because we have $Z \rightarrow X$.
    - Hence, path is blocked by $pa(X)$.
  - Case 2: $Z \notin pa(X)$:
    - Moving downward from $Z$, we can reach a converging node on the path.
    - The converging node and its descendants are not in $pa(X)$.
    - The path is blocked by $pa(X)$. 
Some Notes

- The local Markov property was first mentioned in Lecture 2, when introducing the concept of BN. It is now proved.

- This also explains why we need to make the causal Markov assumption when we causality to build BN structure (slide 36 of Lecture 2):
  - If you use a causal network as a Bayesian network, then we are assuming that causality implies the local Markov property.
Outline

1. An Intuitive Account
   - Special Cases
   - The General Case

2. D-Separation and Independence
   - Some Lemmas
   - Proof of Main Result
   - Corollaries

3. Representing Independence using DAG
A joint distribution $P(\mathbf{V})$ entails conditional independence relationships among variables:

- Use $\mathbf{X} \perp_{P} \mathbf{Y} | \mathbf{Z}$ denotes the fact that, under $P$, $\mathbf{X}$ and $\mathbf{Y}$ are conditional independent given $\mathbf{Z}$, i.e.,

$$P(\mathbf{X}, \mathbf{Y} | \mathbf{Z}) = P(\mathbf{X} | \mathbf{Z})P(\mathbf{Y} | \mathbf{Z})$$

whenever $P(\mathbf{Z}) > 0$

In a DAG $\mathcal{G}$, there D-separation relationships:

- Use $S_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ denotes that the fact that $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$. 
Representing independence using DAG

- \( P(\mathbf{V}) \) obeys the **global Markov property** according to \( G \) if for any three disjoint subsets of variables \( X, Y, \) and \( Z \).

\[
S_G(X,Y,Z) \text{ implies } X \perp_P Y|Z
\]

- When it is the case, we say that \( G \) **represents** some of the independence relationships entailed by \( P \):
  - We can identify independence under \( P \) by examining \( G \).

- When can we use a DAG \( G \) to represent independence relationships entailed by a joint distribution \( P \)?
Factorization

- $P(\mathbf{V})$ factorizes according to $G$ if there exists a Bayesian network such that
  - Its network structure is $G$
  - The joint probability it represents is $P(\mathbf{V})$. 
Local Markov properties

- $P(V)$ obeys the **local Markov property** according to $G$ if for any variable $X$

  $$X \perp_P nd_G(X) | pa_G(X)$$

  where $nd(X)$ stands for the set of non-descendants of $X$. 
Representing Independence using DAG

Factorization and independence

Theorem (3.2)

Let $P(\mathbf{V})$ be a joint probability and $G$ be a DAG over a set of variables $\mathbf{V}$. The following statements are equivalent:

1. $P(\mathbf{V})$ factorizes according to $G$.
2. $P(\mathbf{V})$ obeys the global Markov property according to $G$.
3. $P(\mathbf{V})$ obeys the local Markov property according to $G$.

Proof:

- $1 \Rightarrow 2$: Theorem 3.1.
- $2 \Rightarrow 3$: Corollary 3.2.
Proof of Theorem 3.2 (cont’d)

3 ⇒ 1:

- Induction on the number of nodes.
- Trivially true where there is only one node.
- Suppose true in the case of \( n-1 \) nodes.
- Consider the case of \( n \) nodes.
  - Let \( X \) be a leaf node in \( G \), \( \mathbf{V}' = \mathbf{V} \setminus \{X\} \).
  - By (3), \( X \) is independent of all other nodes given \( \text{pa}(X) \).
  - Hence

\[
P(\mathbf{V}) = P(\mathbf{V}')P(X|\mathbf{V}') = P(\mathbf{V}')P(X|\text{pa}(X))
\]

- Let \( G' \) be obtained from \( G \) by removing \( X \).
- Then \( P(\mathbf{V}') \) obeys the local Markov property according to \( G' \).
- Since there are only \( n-1 \) nodes in \( \mathbf{V}' \), \( P(\mathbf{V}') \) factorizes according to \( G' \).
- Hence \( P(\mathbf{V}) \) factorizes according to \( G \).

- The theorem is proved. Q.E.D
I-Map and D-Map

- $\mathcal{G}$ is an **I-map** of $P(V)$ if for any three disjoint subsets of variables $X$, $Y$, and $Z$:

  $$S_{\mathcal{G}}(X, Y, Z) \text{ implies } X \perp_P Y | Z$$

  i.e. d-Separation in DAG implies independence.

- $\mathcal{G}$ is an **D-map** of $P(V)$ if

  $$X \perp_P Y | Z \text{ implies } S_{\mathcal{G}}(X, Y, Z)$$

  i.e. Independence implies separation in DAG. Non-separation implies dependence.

- $\mathcal{G}$ is an **perfect map** of $P(V)$ if

  - it is both an I-map and a D-map.

  This is ideal case. But there are joint distributions that do not have perfect maps. (Can you think of one?)
I-Map and D-Map

- Adding an edge in an I-map results in another I-map. (Exercise)

- Deleting an edge in a D-map results in another D-Map. (Exercise)

- A **minimal I-map** of $P(V)$ is an I-map such that deletion of one edge will render the graph a non-I-map.

- When constructing BN structure following the procedure given on Slide 24 of Lecture 2,
  - If $pa(X_i)$ is selected to be minimal, then resulting network is an I-map of $P$. 