Exact Inference Algorithms for Probabilistic Reasoning; BTE and CTE

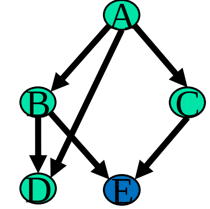
COMPSCI 276, Spring 2017

Set 6: Rina Dechter

(Reading: Primary: Dechter chapter 5
Secondary: , Darwiche chapters 7,8)

Bucket elimination

Algorithm *BE-bel* (Dechter 1996)



$$P(A \mid E=0) = \alpha \sum_{E=0,D,C,B} P(A)\cdot P(B \mid A)\cdot P(C \mid A)\cdot P(D \mid A,B)\cdot P(E \mid B,C)$$

$\sum_b \prod$ ← Elimination operator

bucket B: P(b|a) P(d|b,a) P(e|b,c)

bucket C: P(c|a) $\lambda^B(a,d,c,e)$

bucket D: $\lambda^C(a,d,e)$

bucket E: e=0 $\lambda^D(a,e)$

bucket A: P(a) $\lambda^E(a)$

P(e=0)

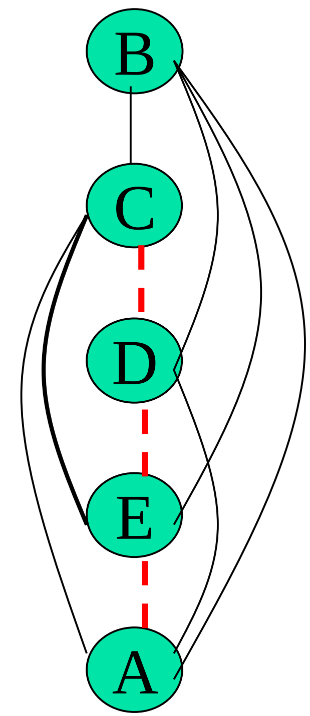
P(a|e=0)

W*=4

"induced width"

(max clique size)

$$P(a \mid e=0) = \frac{P(a,e=0)}{P(e=0)}$$



Finding $MPE = \max_{\bar{x}} P(\bar{x})$

Algorithm *BE-mpe* (Dechter 1996)

$\sum$ is replaced by ***max*** :

$$MPE = \max_{a,e,d,c,b} P(a)P(c \mid a)P(b \mid a)P(d \mid a,b)P(e \mid b,c)$$

$\max_b \prod$ ← Elimination operator

bucket B: P(b|a) P(d|b,a) P(e|b,c)

bucket C: P(c|a) $h^B(a,d,c,e)$

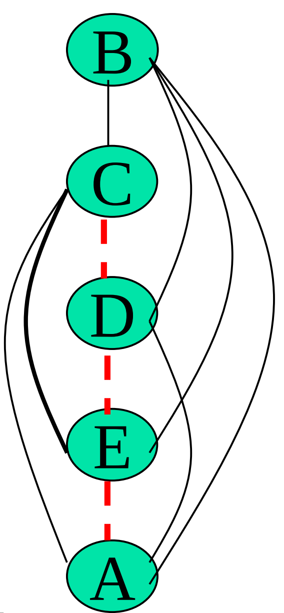
bucket D: $h^C(a,d,e)$

bucket E: e=0 $h^D(a,e)$

bucket A: P(a) $h^E(a)$

MPE

W*=4
"induced width"
(max clique size)



Generating the MPE-tuple

5. $b' = \arg\max_b P(b \mid a') \times P(d' \mid b, a') \times P(e' \mid b, c')$

4. $c' = \arg\max_c P(c \mid a') \times h^B(a', d', c, e')$

3. $d' = \arg\max_d h^C(a', d, e')$

2. $e' = 0$

1. $a' = \arg\max_a P(a) \cdot h^E(a)$

B: $P(b|a)$ $P(d|b,a)$ $P(e|b,c)$

C: $P(c|a)$ $h^B(a,d,c,e)$

D: $h^C(a,d,e)$

E: $e=0$ $h^D(a,e)$

A: $P(a)$ $h^E(a)$

Return (a', b', c', d', e')

Probabilistic Inference Tasks

- Belief updating:

$$\mathbf{BEL(X_i) = P(X_i = x_i \mid evidence)}$$

- Finding most probable explanation (MPE)

$$\bar{x}^* = \arg\max_{\bar{x}} P(\bar{x}, e)$$

- Finding maximum a-posteriory hypothesis

$$(a_1^*, \ldots, a_k^*) = \arg\max_{\bar{a}} \sum_{X/A} P(\bar{x}, e)$$

$A \subseteq X$: **hypothesis variables**

- Finding maximum-expected-utility (MEU) decision

$$(d_1^*, \ldots, d_k^*) = \arg\max_{\bar{d}} \sum_{X/D} P(\bar{x}, e) U(\bar{x})$$

$D \subseteq X$: **decision variables**
$U(\bar{x})$: **utility function**

Markov Networks

Definition 2.23 Markov networks. A Markov network is a graphical model $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{H}, \prod \rangle$ where $\mathbf{H} = \{\psi_1, \ldots, \psi_m\}$ is a set of potential functions where each potential ψ_i is a non-negative real-valued function defined over a scope of variables $\mathcal{S} = \{\mathbf{S}_1, ..., \mathbf{S}_m\}$. $\mathbf{S}_i$. The Markov network represents a global joint distribution over the variables $\mathbf{X}$ given by:

$$P_{\mathcal{M}} = \frac{1}{Z} \prod_{i=1}^{m} \psi_i \quad , \quad Z = \sum_{\mathbf{X}} \prod_{i=1}^{m} \psi_i$$

where the normalizing constant Z is called the partition function.

Everything is applicable to Markov networks as well

Dechter, chapter 2

General Graphical Models

Definition 2.2 Graphical model. A *graphical model* $\mathcal{M}$ is a 4-tuple, $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \otimes \rangle$, where:

1. $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a finite set of variables;
2. $\mathbf{D} = \{D_1, \ldots, D_n\}$ is the set of their respective finite domains of values;
3. $\mathbf{F} = \{f_1, \ldots, f_r\}$ is a set of positive real-valued discrete functions, defined over scopes of variables $\mathcal{S} = \{S_1, ..., S_r\}$, where $\mathbf{S}_i \subseteq \mathbf{X}$. They are called *local* functions.
4. $\otimes$ is a *combination* operator (e.g., $\otimes \in \{\prod, \sum, \bowtie\}$ (product, sum, join)). The combination operator can also be defined axiomatically as in [Shenoy, 1992], but for the sake of our discussion we can define it explicitly, by enumeration.

The graphical model represents a *global function* whose scope is $\mathbf{X}$ which is the combination of all its functions: $\bigotimes_{i=1}^{r} f_i$.

General Bucket Elimination

Algorithm General bucket elimination (GBE)

Input: $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \otimes \rangle$. $F = \{f_1, ..., f_n\}$ an ordering of the variables, $d = X_1, ..., X_n$; $\mathbf{Y} \subseteq \mathbf{X}$.

Output: A new compiled set of functions from which the query $\Downarrow_Y \otimes_{i=1}^{n} f_i$ can be derived in linear time.

1. **Initialize:** Generate an ordered partition of the functions into $bucket_1, ..., bucket_n$, where $bucket_i$ contains all the functions whose highest variable in their scope is X_i. An input function in each bucket ψ_i, $\psi_i = \otimes_{i=1}^{n} f_i$.
2. **Backward:** For $p \leftarrow n$ downto 1, do
for all the functions $\psi_p, \lambda_1, \lambda_2, ..., \lambda_j$ in $bucket_p$, do

- **If** (observed variable) $X_p = x_p$ appears in $bucket_p$, assign $X_p = x_p$ in ψ_p and to each λ_i and put each resulting function in appropriate bucket.
- **else,** (combine and marginalize)
$\lambda_p \leftarrow \Downarrow_{S_p} \psi_p \otimes (\otimes_{i=1}^{j} \lambda_i)$ and add λ_p to the largest-index variable in $scope(\lambda_p)$.

3. **Return:** all the functions in each bucket.

Theorem 4.23 Correctness and complexity. *Algorithm GBE is sound and complete for its task. Its time and space complexities is exponential in the* $w^*(d) + 1$ *and* $w^*(d)$, *respectively, along the order of processing* d.

1998 roadmap

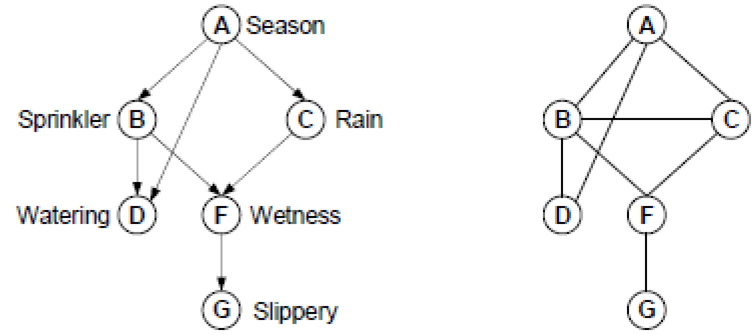
Outline; Road Map

Methods \ Tasks	CSP	SAT	Optimization	Belief updating	MPE, MAP, MEU	Solving linear equalities/ inequalities
elimination	adaptive consistency join-tree	directional resolution	dynamic programming	join-tree, VE, SPI, elim-bel	join-tree, elim-mpe, elim-map	Gaussian/ Fourier elimination
conditioning	backtracking search	backtracking (Davis-Putnam)	branch-and-bound, best-first search		branch-and-bound, best-first search	
elimination + conditioning	cycle-cutset forward checking	DCDR, BDR-DP		loop-cutset		
approximate elimination	i-consistency	bounded (directional) resolution	mini-buckets	mini-buckets	mini-buckets	
approximate conditioning	greedylocal search (GSAT)	GSAT	gradient descent	stochastic simulation	gradient descent	
approximate (elimination + conditioning)	GSAT + partial path-consistency					

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

A Bayesian Network Processed by BE



(a) Directed acyclic graph (b) Moral graph

Two orderings for BE

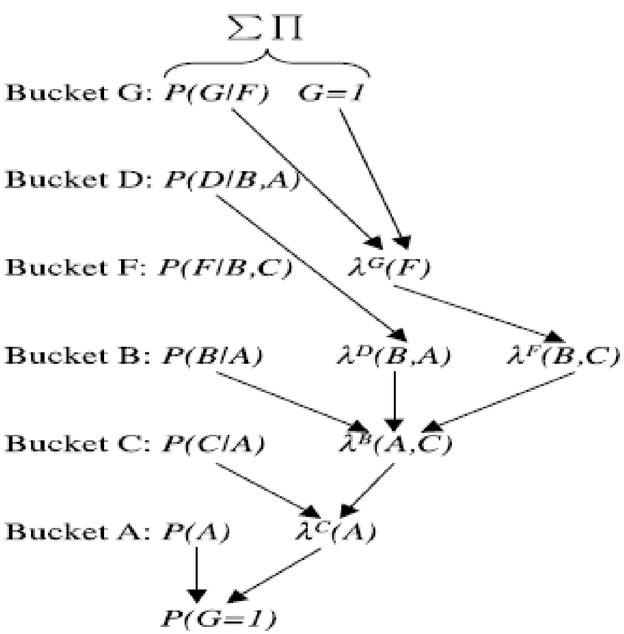
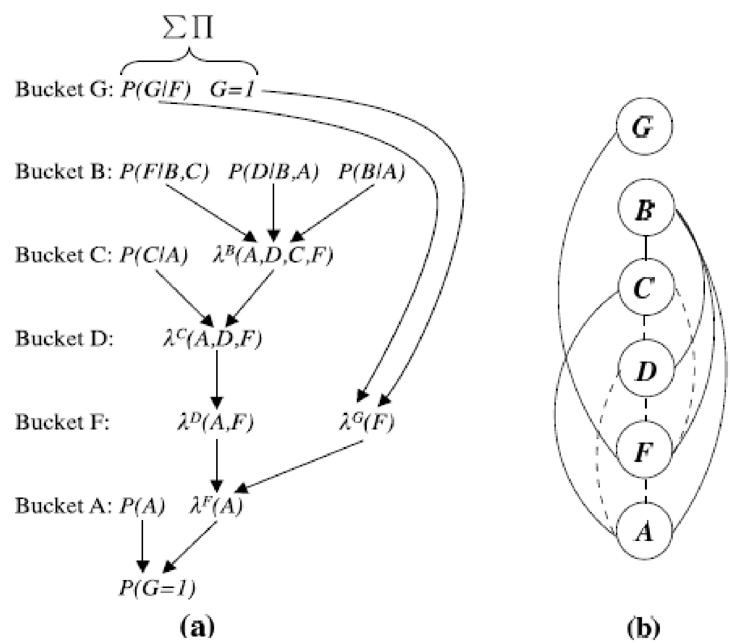


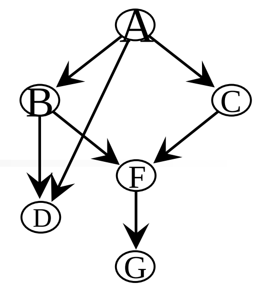
Figure 4.2: Bucket elimination along ordering $d_1 = A, C, B, F, D, G$.



.3: The bucket's output when processing along $d_2 = A, F, D, C, B,$

Complexity exponential in $w^*(d)$

From Bucket Elimination to Bucket-Tree Elimination



Bucket G: $P(G|F)$

Bucket F: $P(F|B,C)$ $\lambda_{G \to F}(F)$

Bucket D: $P(D|A,B)$

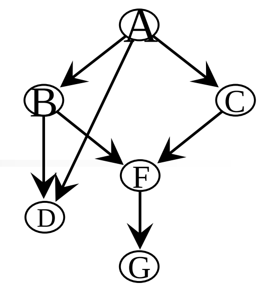
Bucket C: $P(C|A)$ $\lambda_{F \to C}(B,C)$

Bucket B: $P(B|A)$ $\lambda_{D \to B}(A,B)$ $\lambda_{C \to B}(A,B)$

Bucket A: $P(A)$ $\lambda_{B \to A}(A)$

Observation 1: BE is a message propagation down a bucket-tree

From Bucket Elimination to Bucket-Tree Elimination



What If we want the marginal on B?

Bucket G: $P(G|F)$

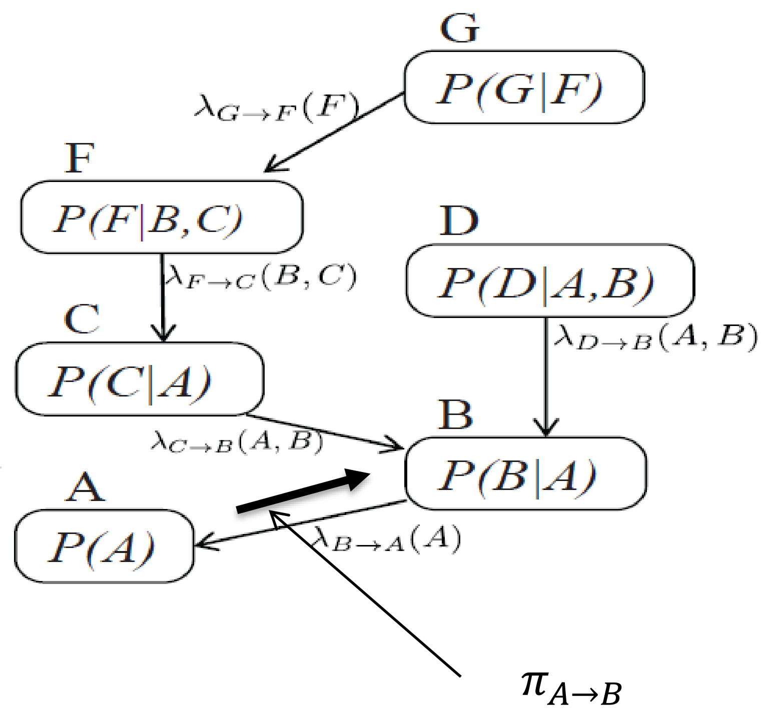
Bucket F: $P(F|B,C)$ $\lambda_{G\to F}(F)$

Bucket D: $P(D|A,B)$

Bucket C: $P(C|A)$ $\lambda_{F\to C}(B,C)$

Bucket B: $P(B|A)$ $\lambda_{D\to B}(A,B)$ $\lambda_{C\to B}(A,B)$

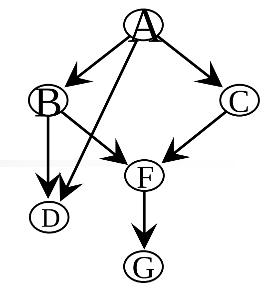
Bucket A: $P(A)$ $\lambda_{B\to A}(A)$



$$\pi_{A\to B}(a) = P(A),$$

$$P(B) = \sum P(B \mid A) P(A) \lambda_{C->B}(B,C) \lambda_{D->B}(A,B)$$

From Bucket Elimination to Bucket-Tree Elimination (BTE)



What if we want the marginal on D?
Imagine combining B and A, D
d = ({A,D,B},C,F,G)

Bucket G: $P(G|F)$

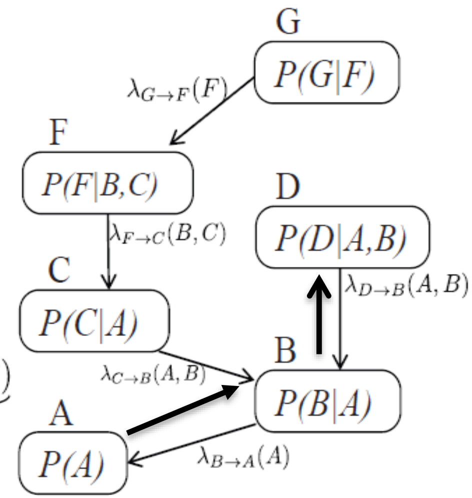
Bucket F: $P(F|B,C)$ $\lambda_{G\to F}(F)$

Bucket D: $P(D|A,B)$

Bucket C: $P(C|A)$ $\lambda_{F\to C}(B,C)$

Bucket B: $P(B|A)$ $\lambda_{D\to B}(A,B)$ $\lambda_{C\to B}(A,B)$

Bucket A: $P(A)$ $\lambda_{B\to A}(A)$



$$\pi_{A\to B}(a) = P(A),$$

$$\pi_{B\to D}(a,b) = p(b|a)\cdot\pi_{A\to B}(a)\cdot\lambda_{C\to B}(b)$$

$$bel(d) = \alpha\sum_{a,b} P(d|a,b)\cdot\pi_{B\to D}(a,b).$$

Idea of BTE

This example generalize: We can compute the belief in each variable by a second message-passing along the bucket-tree.

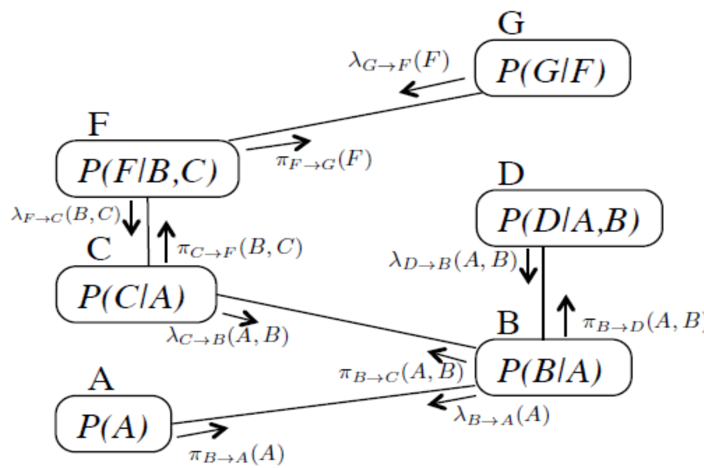
in Bayesian networks. Given an ordering of the variables d the first step generates the bucket-tree by partitioning the functions into buckets and connecting the buckets into a tree. The subsequent *top-down* phase is identical to general bucket-elimination. The *bottom-up* messages are defined as follows. The messages sent from the root up to the leaves will be denoted by π. The message from B_j to a child B_i is generated by combining (e.g., multiplying) all the functions currently in B_j including the π messages from its parent bucket and all the λ messages from its *other* child buckets and marginalizing (e.g., summing) over the eliminator from B_j to B_i. By construction, downward messages are generated by eliminating a single variable. Upward messages, on the other hand, may be generated by eliminating zero, one or more variables.

Idea of BTE

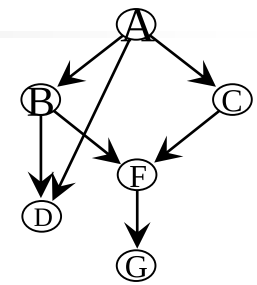
This example generalize: We can compute the belief in each variable by a second message-passing along the bucket-tree.

BTE: Allows Messages Both Ways

Initial buckets
+ messages



(a)



$$\pi_{A\to B}(a) = P(a)$$

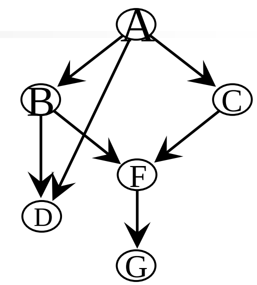
$$\pi_{B\to C}(c,a) = P(b|a)\lambda_{D\to B}(a,b)\pi_{A\to B}(a)$$

$$\pi_{B\to D}(a,b) = P(b|a)\lambda_{C\to B}(a,b)\pi_{A\to B}(a,b)$$

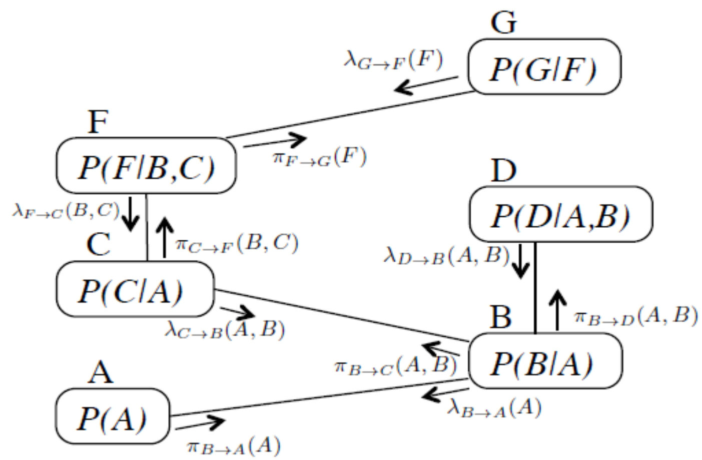
$$\pi_{C\to F}(c,b) = \sum_a P(c|a)\pi_{B\to C}(a,b)$$

$$\pi_{F\to G}(f) = \sum_{b,c} P(f|b,c)\pi_{C\to F}(c,b)$$

BTE: Allows Messages Both Ways

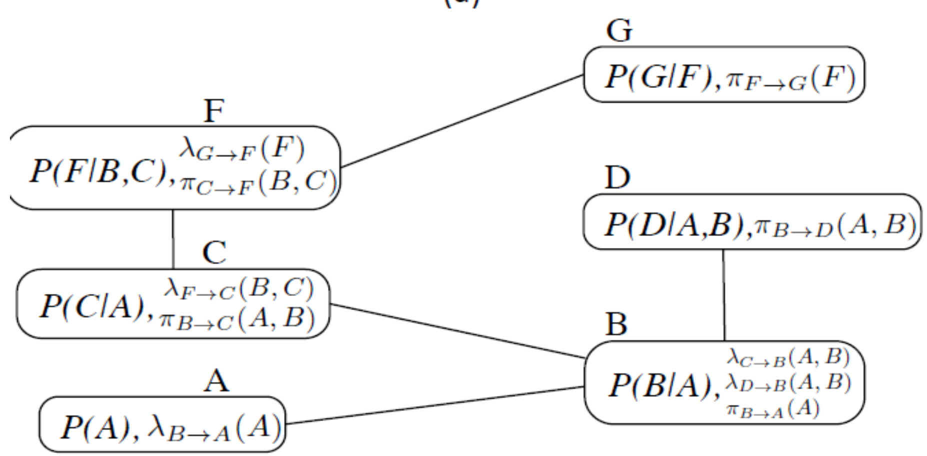


Initial buckets
+ messages



(a)

Output buckets



(b)

A Bucket-Tree of a Bayesian Network

- The bucket-tree:
- Nodes are the buckets. Each has functions (assigned initially) and variables: itself+ induced-parents
  - There is an arc from B_i to B_j iff the function created at bucket B_i is placed at bucket B_j
  - We have a separator and eliminator between two adjacent buckets

Bucket-Tree Construction From the Graph

1. Pick a (good) variable ordering, d.
2. Generate the induced ordered graph
3. From top to bottom, each bucket of X is mapped to (variables,functions) pairs
4. The variables are the clique of X, the functions are those placed in the bucket
5. Connect the bucket of X to earlier bucket of Y if Y is the closest node connected to X

BTE

Theorem: When BTE terminates The product of functions in each bucket is the beliefs of the variables joint with the evidence.

ALGORITHM BUCKET-TREE ELIMINATION (BTE)

Input: A problem $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \prod \rangle$, ordering d. $X = \{X_1, ..., X_n\}$ and $F = \{f_1, ..., f_r\}$ Evidence $E = e$.

Output: Augmented buckets $\{B'_i\}$, containing the original functions and all the π and λ functions received from neighbors in the bucket-tree.

1. **Pre-processing:** Partition functions to the ordered buckets as usual and generate the bucket-tree.
2. **Top-down phase:** λ messages (BE) **do**

   for $i = n$ to 1, in reverse order of d process bucket B_i:

   The message $\lambda_{i \to j}$ from B_i to its parent B_j, is:

   $\lambda_{i \to j} \Leftarrow \sum_{elim(i,j)} \psi_i \cdot \prod_{k \in child(i))} \lambda_{k \to i}$

   endfor
3. **bottom-up phase:** π **messages**

   for $j = 1$ to n, process bucket B_j **do**:

   B_j takes $\pi_{k \to j}$ received from its parent B_k, and computes a message $\pi_{j \to i}$ for each child bucket B_i by

   $\pi_{j \to i} \Leftarrow \sum_{elim(j,i)} \pi_{k \to j} \cdot \psi_j \cdot \prod_{r \neq i} \lambda_{r \to j}$

   endfor
4. **Output:** and answering singleton queries (e.g., deriving beliefs).

   Output augmented buckets $B'_1, ..., B'_n$, where each B'_i contains the original bucket functions and the λ and π messages it received.

Query Answering

COMPUTING MARGINAL BELIEFS

Input: a bucket tree processed by BTE with augmented buckets: $B'_1, ..., B'_n$

output: beliefs of each variable, bucket, and probability of evidence.

$$bel(B_i) \Leftarrow \alpha \cdot \prod_{f \in B'_i} f$$

$$bel(X_i) \Leftarrow \alpha \cdot \sum_{B_i - \{X_i\}} \prod_{f \in B'_i} f$$

$$P(evidence) \Leftarrow \sum_{B_i} \prod_{f \in B'_i} f$$

Figure 5.4: Query answering.

Explicit functions

Definition 5.4 Explicit function and explicit sub-model. Given a graphical model $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \prod \rangle$, and reasoning tasks defined by marginalization $\sum$ and given a subset of variables Y, $Y \subseteq \mathbf{X}$, we define $\mathcal{M}_Y$, the explicit function of $\mathcal{M}$ over Y:

$$\mathcal{M}_Y = \sum_{X-Y} \prod_{f \in F} f, \tag{5.4}$$

We denote by F_Y any set of functions whose scopes are subsumed in Y over the same domains and ranges as the functions in $\mathbf{F}$. We say that (Y, F_Y) is an explicit submodel of $\mathcal{M}$ iff

$$\prod_{f \in F_Y} f = \mathcal{M}_Y \tag{5.5}$$

Asynchronous BTE: Bucket-tree Propagation (BTP)

BUCKET-TREE PROPAGATION (BTP)

Input: A problem $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \prod, \sum \rangle$, ordering d. $X = \{X_1, ..., X_n\}$ and $F = \{f_1, ..., f_r\}$, $\mathbf{E} = \mathbf{e}$. An ordering d and a corresponding bucket-tree structure, in which for each node X_i, its bucket B_i and its neighboring buckets are well defined.

Output: Explicit buckets. Assume functions assigned with the evidence.

1. **for** bucket B_i **do:**
2. **for** each neighbor bucket B_j **do**,
   once all messages from all other neighbors were received, **do**
   compute and send to B_j the message
   $$\lambda_{i \to j} \Leftarrow \sum_{elim(i,j)} \psi_i \cdot \left(\prod_{k \neq j} \lambda_{k \to i}\right)$$
3. **Output:** augmented buckets $B'_1, ..., B'_n$, where each B'_i contains the original bucket functions and the λ messages it received.

Properties of BTE

- Theorem (correctness) 5.5 *Algorithm BTE when applied to a Bayesian or Markov network is sound. Namely, in each bucket we can exactly compute the exact joint function of every subset of variables and the evidence.*
- *(follows from imapness of trees)*

•Theorem 5.6 (complexity of BTE) *Let w*(d) be the induced width of G along ordering d, r be the number of functions and k the maximum domain size. BTE is* $O(r \cdot deg \cdot k^{(w*(d)+1)})$ *time, where deg is the maximum degree in the bucket-tree. BTE is* $O(n\ k^{w^*(d)}.)$ *space*

Complexity of BTE/BTP on Trees

Theorem 5.6 Complexity of BTE. *Let $w^*(d)$ be the induced width of (G^*, d) where G is the primal graph of $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \prod, \sum \rangle$, r be the number of functions in $\mathbf{F}$ and k be the maximum domain size. The time complexity of BTE is $O(r \cdot deg \cdot k^{w^*(d)+1})$, where deg is the maximum degree of a node in the bucket tree. The space complexity of BTE is $O(n \cdot k^{w^*(d)})$.*

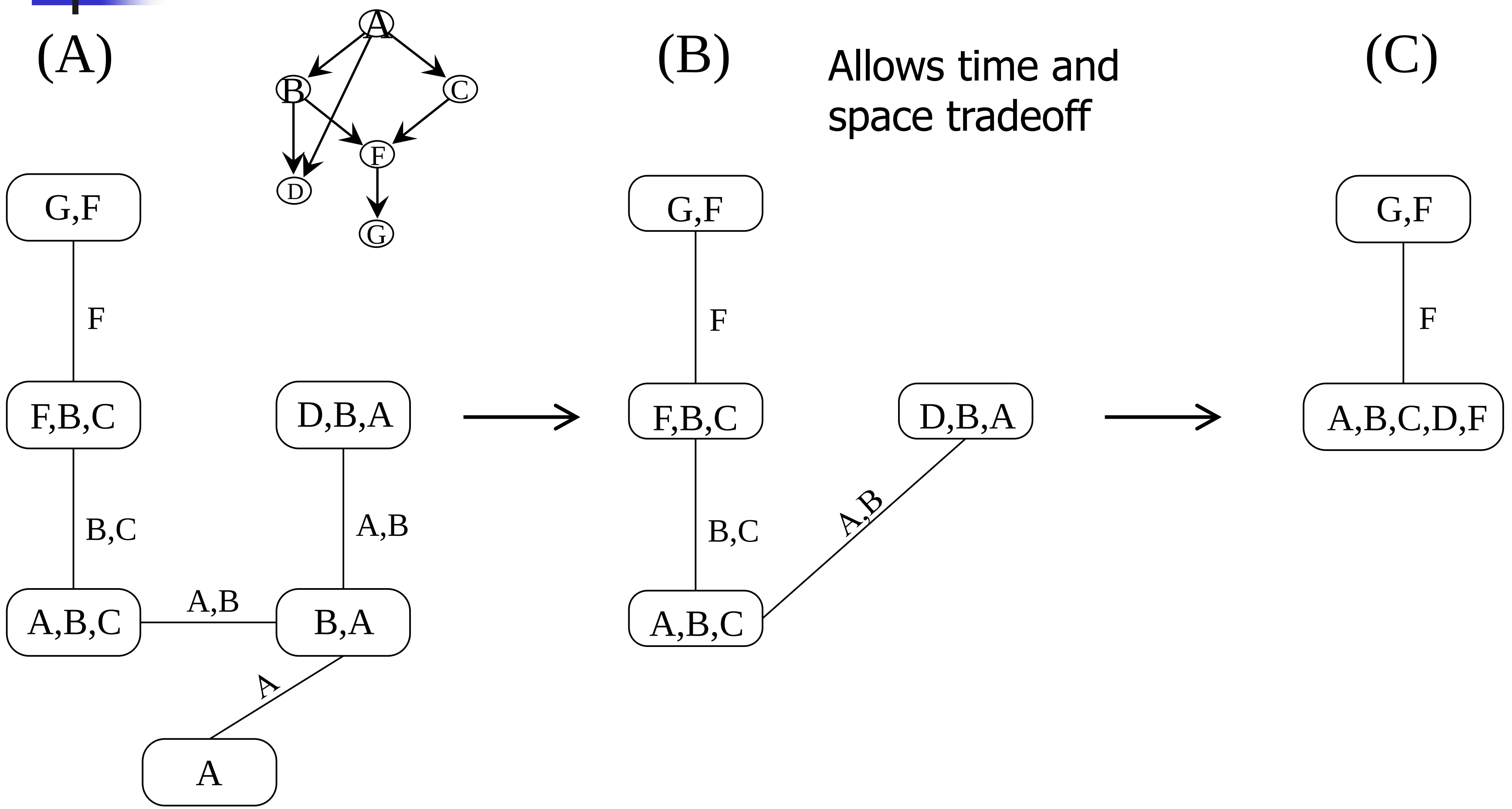
Proposition 5.8 BTE on trees *For tree graphical models, algorithms BTE and BTP are time and space $O(nk^2)$ and $O(nk)$, respectively, when k bound the domain size and n bounds the number of variables.*

This will be extended to acyclic graphical models shortly

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

From Buckets to Clusters

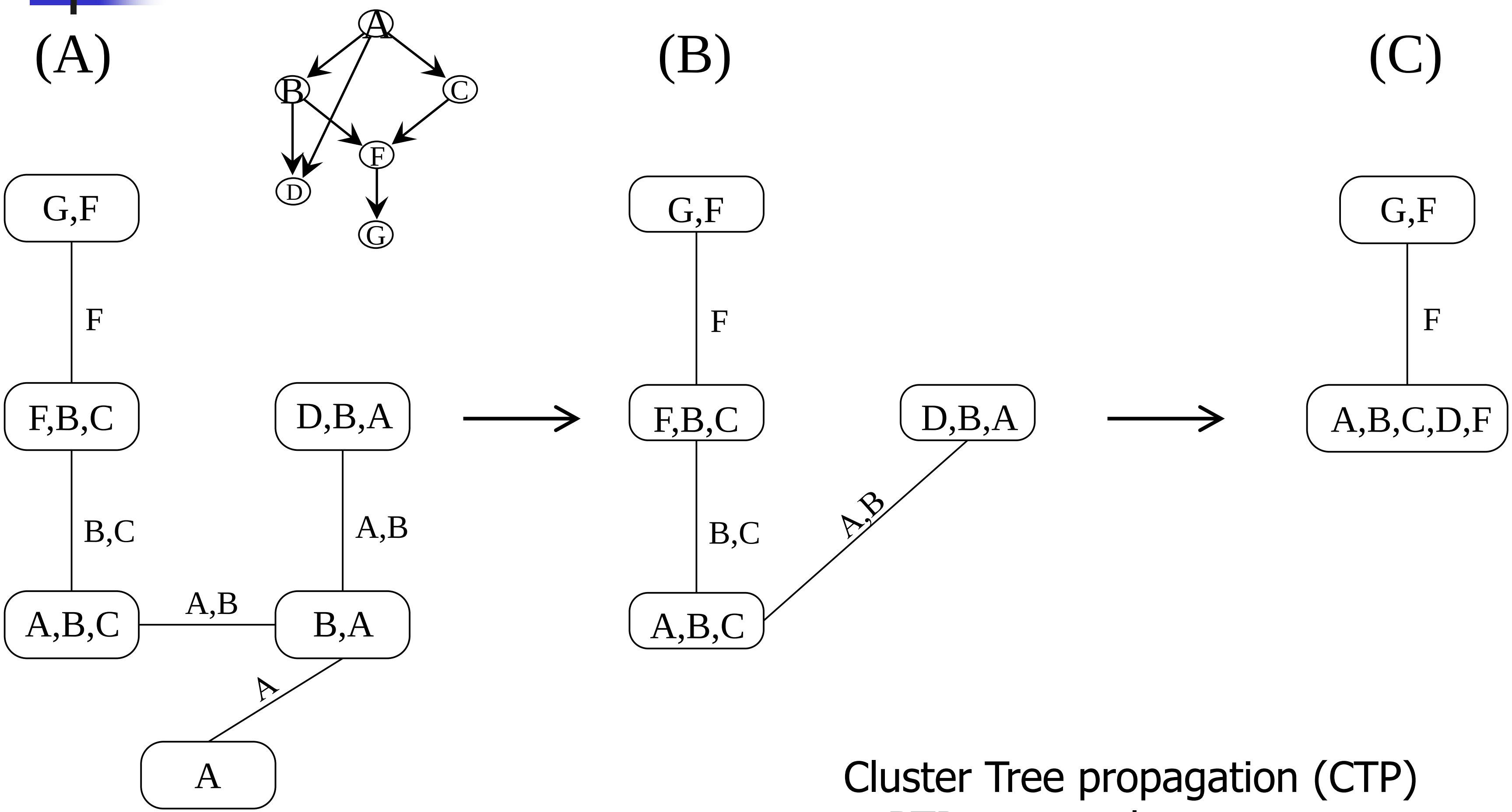


Allows time and space tradeoff

From a Bucket-Tree to a Cluster-Tree

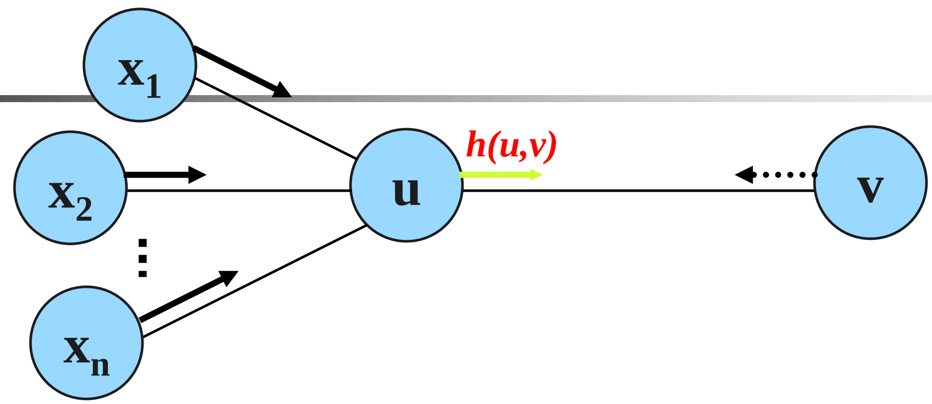
- Merge none-maximal buckets into maximal clusters.
- Connect clusters into a tree: each cluster to one with which it shares a largest subset of variables.
- Separators are the intersection of variables on the arcs of the tree.
- The cluster-tree is an i-map.

Examples



Cluster Tree propagation (CTP)
= BTP on any cluster-tree.

The General Message Passing On a General Tree-Decomposition (CTP)



$$cluster(u) = \psi(u) \cup \{h(x_1,u), h(x_2,u), \ldots, h(x_n,u), h(v,u)\}$$

For max-product
Just replace $\sum$
With max.

Compute the message :

$$h(u,v) = \sum_{elim(u,v)} \prod_{f \in cluster(u) - \{h(v,u)\}} f$$

$h_{u \to v}$ *new notation*
for h(u,v)

elim(u,v) = cluster(u)-sep(u,v)

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

Acyclic Networks

- **Dual network**: Each scope of a CPT is a node and each arc is denoted by intersection.
- **Acylic network:** when the dual graph is a tree or has a join-tree
- **Acyclic network (alternative characteristic)**: A network is acyclic if it has a tree-decomposition where each node has a single original CPT.
- Tree-clustering converts a network into an acyclic one.

Sometime the dual graph seems to not be a tree, but it is in fact, a tree. This is because some of its arcs are redundant and can be removed while not violating the original independency relationships that is captured by the graph.

Acyclic Networks

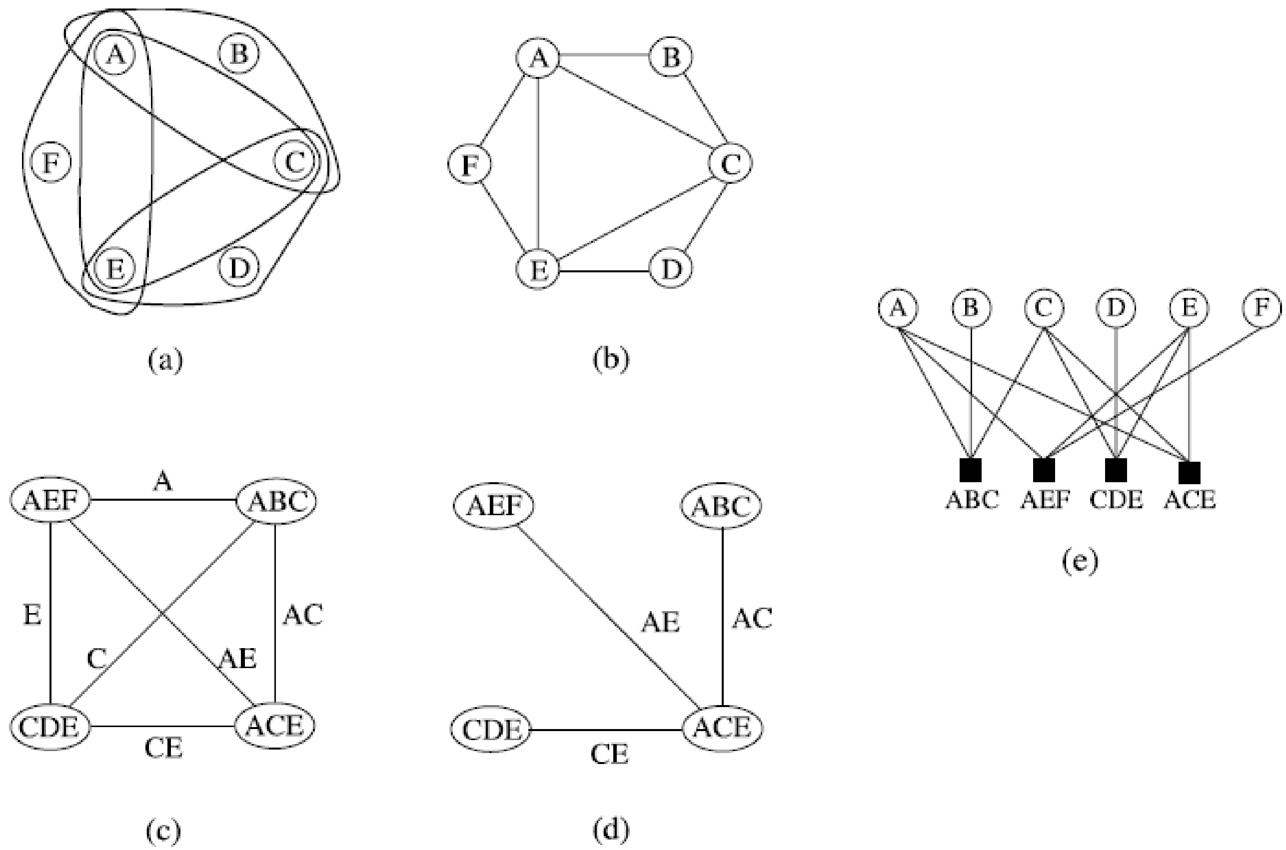


Figure 5.1: (a)Hyper, (b)Primal, (c)Dual and (d)Join-tree of a graphical model having scopes ABC, AEF, CDE and ACE. (e) the factor graph

Connectedness and Join-trees

Definition 5.11 Connectedness, join-trees. Given a dual graph of a graphical model $\mathcal{M}$, an arc subgraph of the dual graph satisfies the *connectedness* property iff for each two nodes that share a variable, there is at least one path of labeled arcs of the dual graph such that each contains the shared variables. An arc subgraph of the dual graph that satisfies the connectedness property is called a *join-graph* and if it is a tree, it is called a *join-tree*.

Also called the running intersection property

Definition: A graphical model whose dual graph has a join-tree is acyclic

Theorem: BTE is time and space linear on acyclic graphical models

Tree-decomposition: If we transform a general model into an acyclic one it can then be solved by a BTE/BTP scheme. Alsp knon as tree-clustering

A Dual Graph Having a Join-Tree is Acyclic

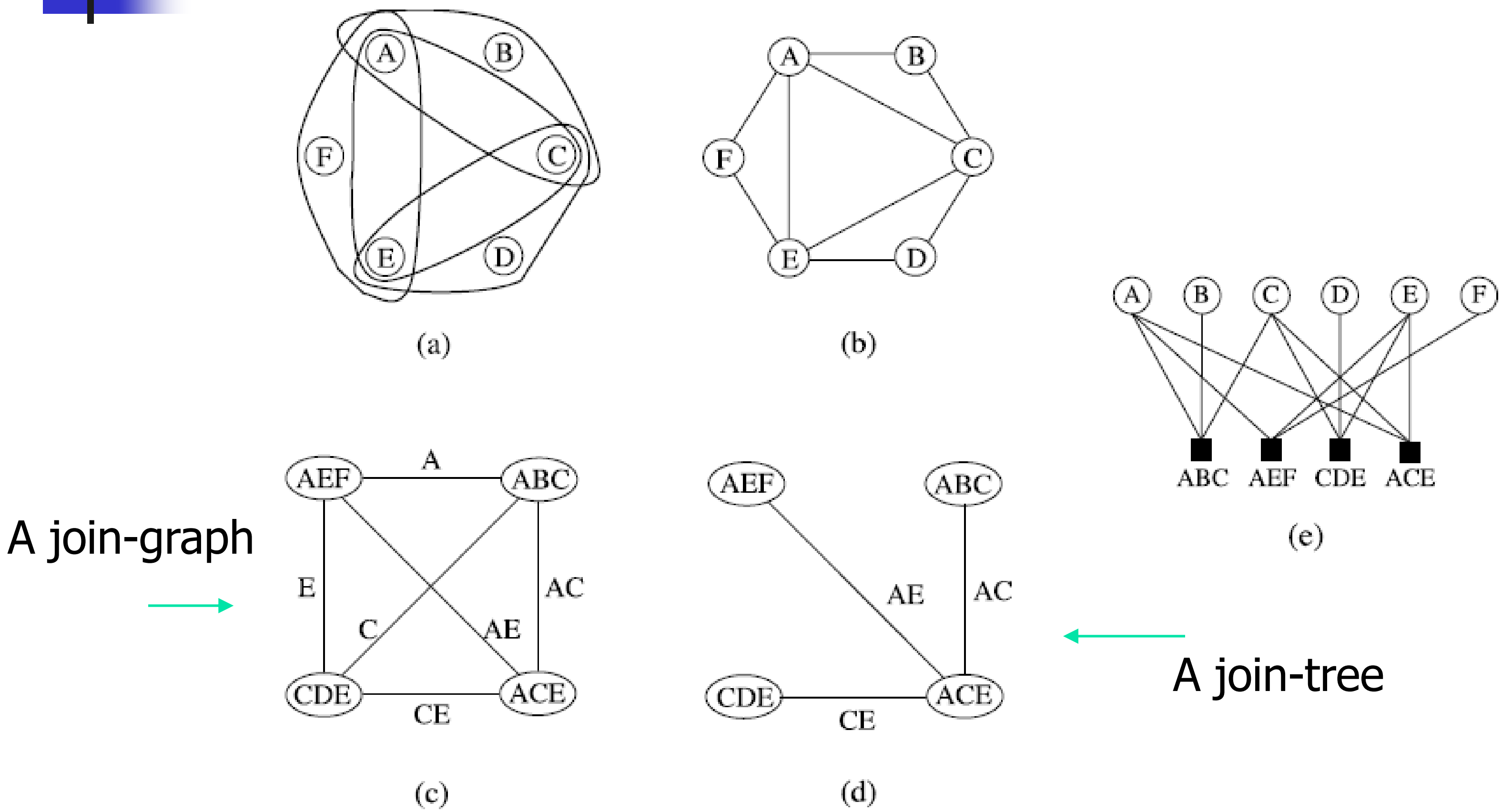


Figure 5.1: (a)Hyper, (b)Primal, (c)Dual and (d)Join-tree of a graphical model having scopes ABC, AEF, CDE and ACE. (e) the factor graph

Tree-decompositions

A *tree decomposition* for a belief network $BN = < X,D,G,P >$ is a triple $< T, \chi, \psi >$, where $T = (V,E)$ is a tree and χ and ψ are labeling functions, associating with each vertex $v \in V$ two sets, $\chi(v) \subseteq X$ and $\psi(v) \subseteq P$ satisfying :

1. For each function $p_i \in P$ there is exactly one vertex such that $p_i \in \psi(v)$ and $scope(p_i) \subseteq \chi(v)$
2. For each variable $X_i \in X$ the set $\{v \in V | X_i \in \chi(v)\}$ forms a connected subtree (running intersection property)

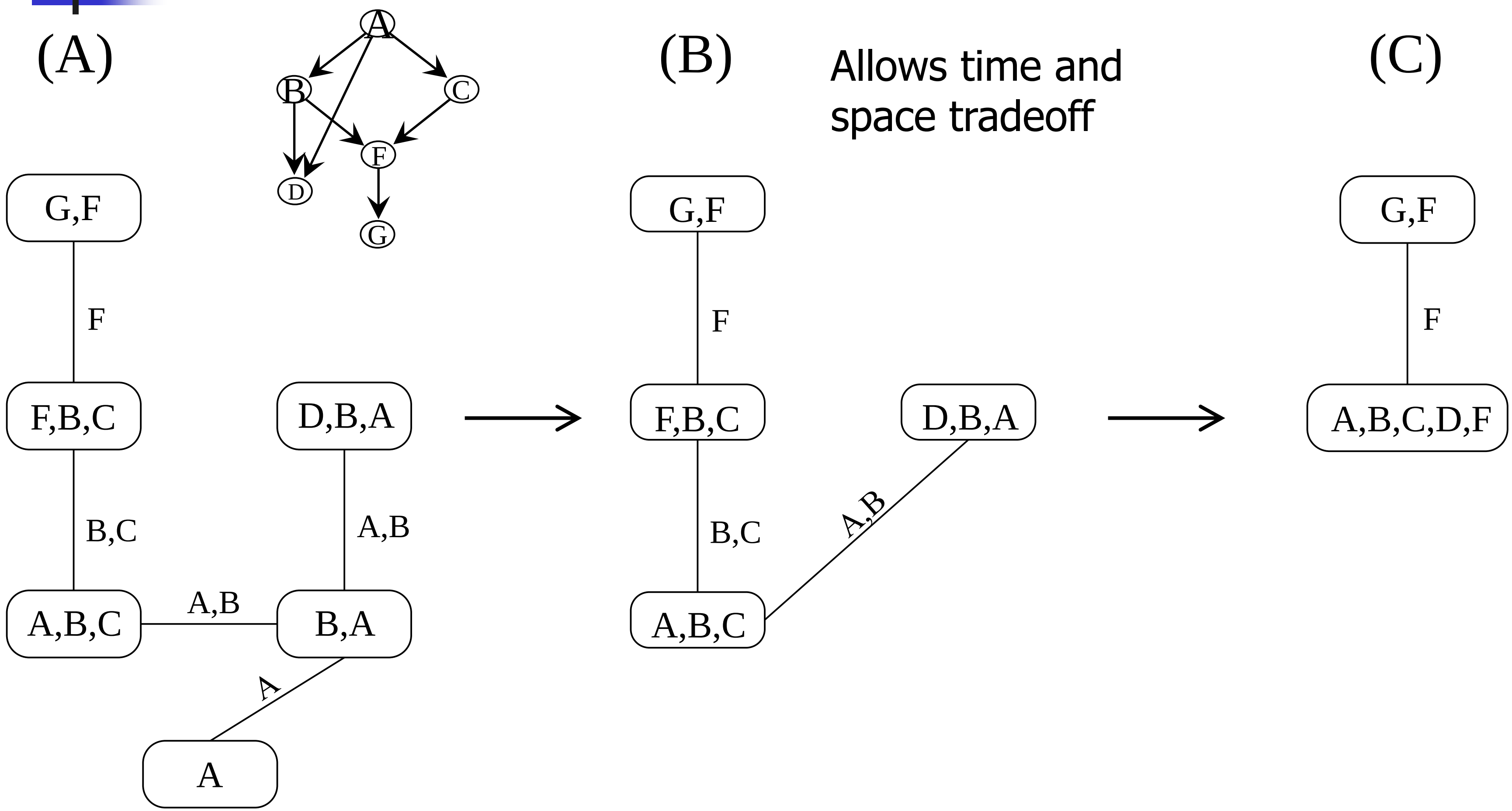
Treewidth: maximum number of nodes in a node of Tree-decomposition – 1
Seperator-width: maximum intersection between adjacent nodes
Eliminator: elim(u,v) = χ(u) - χ(v)

Generating Tree-Decomposition

Proposition 6.2.12 *If T is a tree-decomposition, then any tree obtained by merging adjacent clusters is also a tree-decomposition.*

A bucket-tree of a graphical model is a tree-decomposition of the model

Examples of Tree-Decompositions or Cluster-Trees



Allows time and space tradeoff

Cluster-Tree Elimination

CLUSTER-TREE ELIMINATION (CTE)

Input: A tree decomposition $< T, \chi, \psi >$ for a problem $M =< X, D, F, \prod, \sum\} >$, $X = \{X_1, ..., X_n\}$, $F = \{f_1, ..., f_r\}$. Evidence $E = e$, $\psi_u = \prod_{f \in \psi(u)} f$

Output: An augmented tree decomposition whose clusters are all model explicit. Namely, a decomposition $< T, \chi, \bar{\psi} >$ where $u \in T$, $\bar{\psi}(u)$ is model explicit relative to $\chi(u)$.

1. **Initialize.** (denote by $m_{u \to v}$ the message sent from vertex u to vertex v.)
2. **Compute messages:**

    For every node u in T, once u received messages from all neighbors but v,

    Process observed variables:

    For each node $u \in T$ assign relevant evidence to $\psi(u)$

    Compute the message:

    $$m_{u \to v} \leftarrow \sum_{\chi(u)-sep(u,v)} \psi_u \cdot \prod_{r \in neighbor(u), r \neq v} m_{r \to u}$$

    endfor

    Note: functions whose scopes do not contain any separator variable do not need to be combined and can be directly passed on to the receiving vertex.
3. **Return:** The explicit tree $< T, \chi, \bar{\psi} >$, where

    $\bar{\psi}(v) \Leftarrow \psi(v) \cup_{u \in neighbor(v)} \{m_{u \to v}\}$

    return the explicit function: for each v, $M_{\chi(v)} = \prod_{f \in \bar{\psi}(v)} f$

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

Chordal Graphs and Join-Trees

- A graph is chordal if every cycle of length at least 4 has a chord
- Finding w* of chordal graph is easy using the **max-cardinality ordering**
- A graph along a max-cardinality order has no fill-in edges iff it is chordal.
- Theorem: The maximal cliques of a chordal have a join-tree. Namely, there is a tree connecting the maximal cliques such that connectedness is obeyed.

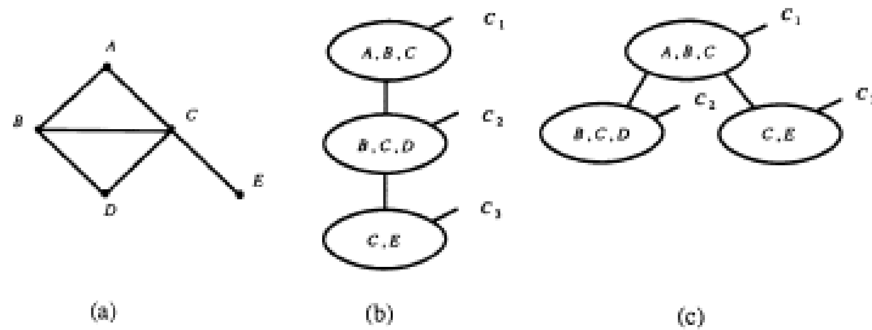
ASSEMBLING A JOIN TREE

1. Use the fill-in algorithm to generate a chordal graph G' (if G is chordal, $G = G'$).

2. Identify all cliques in G'. Since any vertex and its parent set (lower ranked nodes connected to it) form a clique in G', the maximum number of cliques is $|V|$.

3. Order the cliques $C_1, C_2, \ldots, C_t$ by rank of the highest vertex in each clique.

4. Form the join tree by connecting each C_i to a predecessor C_j $(j < i)$ sharing the highest number of vertices with C_i.



EXAMPLE: Consider the graph in Figure 3.9*a*. One maximum cardinality ordering is (A, B, C, D, E).

- Every vertex in this ordering has its preceding neighbors already connected, hence the graph is chordal and no edges need be added.
- The cliques are ranked C_1, C_2, and C_3 as shown in Figure 3.9*b*.
- $C_3 = \{C, E\}$ shares only vertex C with its predecessors C_2 and C_1, so either one can be chosen as the parent of C_3.
- These two choices yield the join trees of Figures 3.9*b* and 3.9*c*.
- Now suppose we wish to assemble a join tree for the same graph with the edge (B, C) missing.
- The ordering (A, B, C, D, E) is still a maximum cardinality ordering, but now when we discover that the preceeding neighbors of node D (i.e., B and C) are nonadjacent, we should fill in edge (B, C).
- This renders the graph chordal, and the rest of the procedure yields the same join trees as in Figures 3.9*b* and 3.9*c*.

Join-Tree Clustering

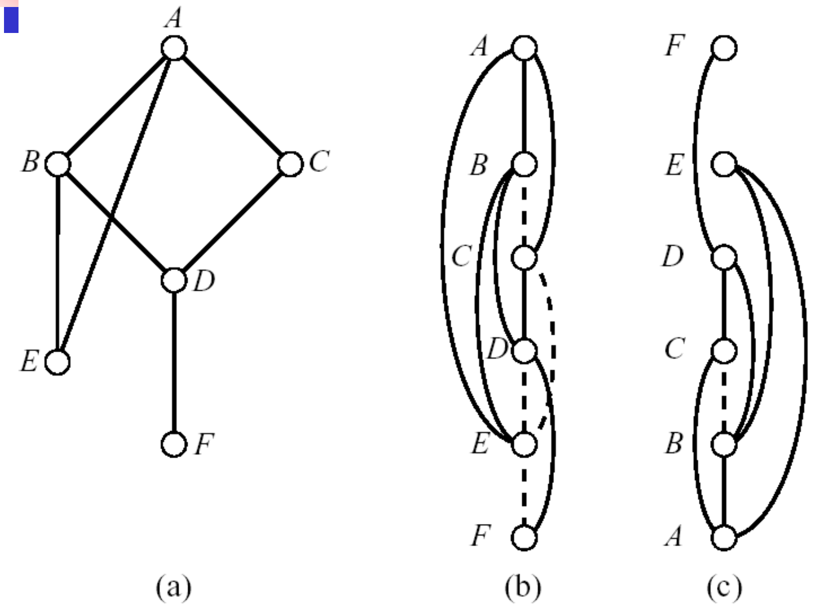
JOIN-TREE CLUSTERING (JTC)

Input: A graphical model $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F}, \prod \rangle$, $\mathbf{X} = \{X_1, ..., X_n\}$, $\mathbf{F} = \{f_1, ..., f_r\}$.
Its scopes $S = S_1, ..., S_r$ and its primal graph is $G = (X, E)$.
Output: A join-tree decomposition $< T, \chi, \psi >$ for $\mathcal{M}$

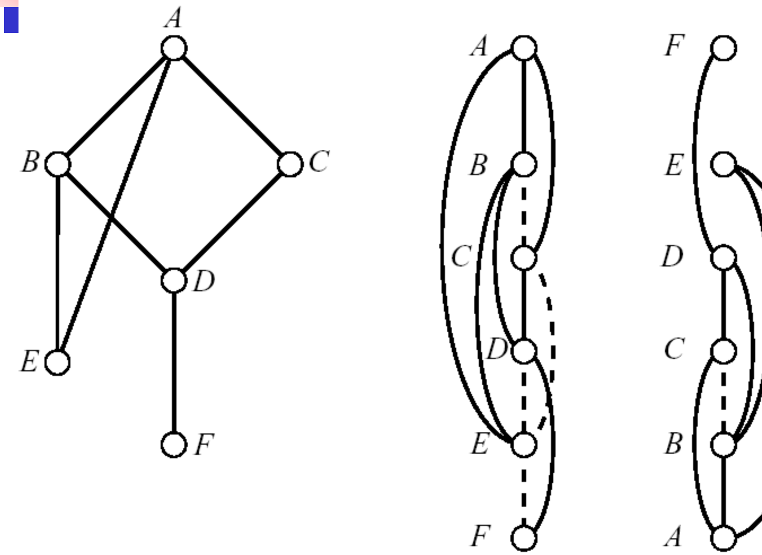
1. Select a variable ordering, $d = (X_1, ..., X_n)$.
2. **Triangulation** (create the induced graph along d and call it G^*):
   for $j = n$ to 1 by -1 do
   $E \leftarrow E \cup \{(i,k) | i < j, \ k < j, \ (i,j) \in E, (k,j) \in E\}$
3. **Create a join-tree of the induced graph** (G^*, d) as follows:
   a. Identify all maximal cliques in the chordal graph.
   Let $C = \{C_1, ..., C_t\}$ be all such cliques, where C_i is the cluster of bucket i.
   b. Create a tree T of cliques:
   Connect each C_i to a C_j $(j < i)$ with whom it shares largest subset of variables.
4. Create ψ_i: Partition input function in cluster-node whose variables contain its scope.
5. Return a tree-decomposition $< T, \chi, \psi >$, where T is generated in step 3,
$\chi(i) = C_i$ and $\psi(i)$ is determined in step 4.

Figure 5.12: Join-tree clustering.

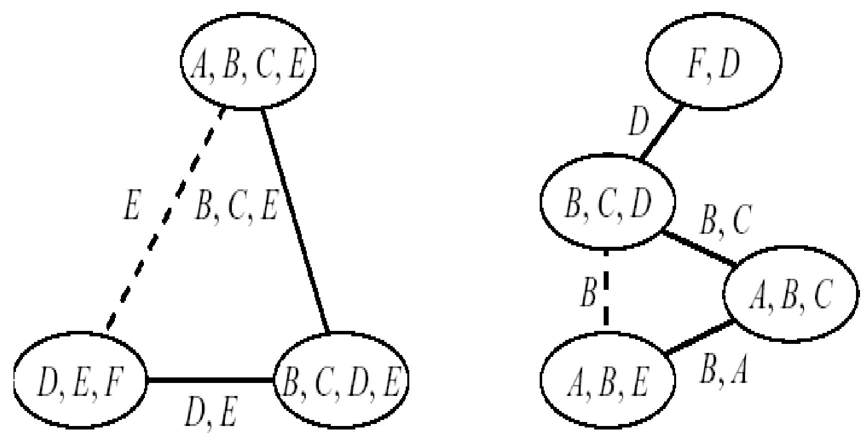
Examples of Generating Join-Trees



Examples of Generating Join-Trees

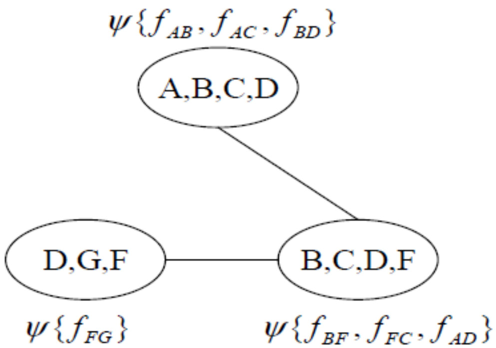


(a) (b) (c)

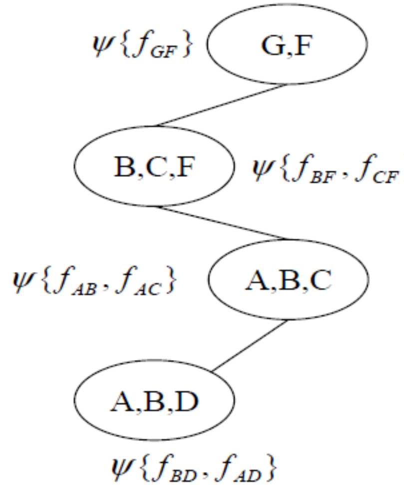


(a) (b)

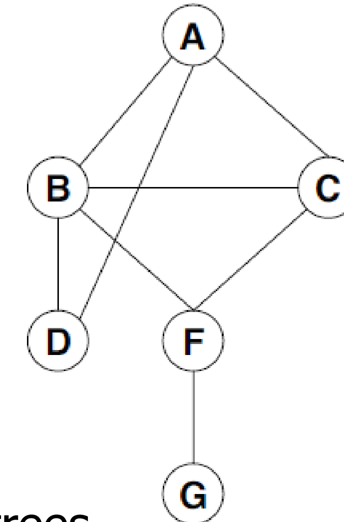
Tree-clustering and message-passing



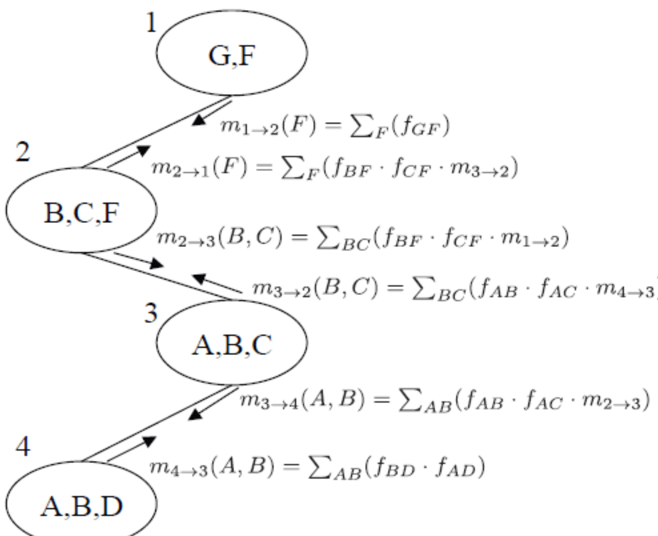
(a)



(b)



Two join-trees



Message-passing by CTE on
The tree in (b)

Properties of CTE

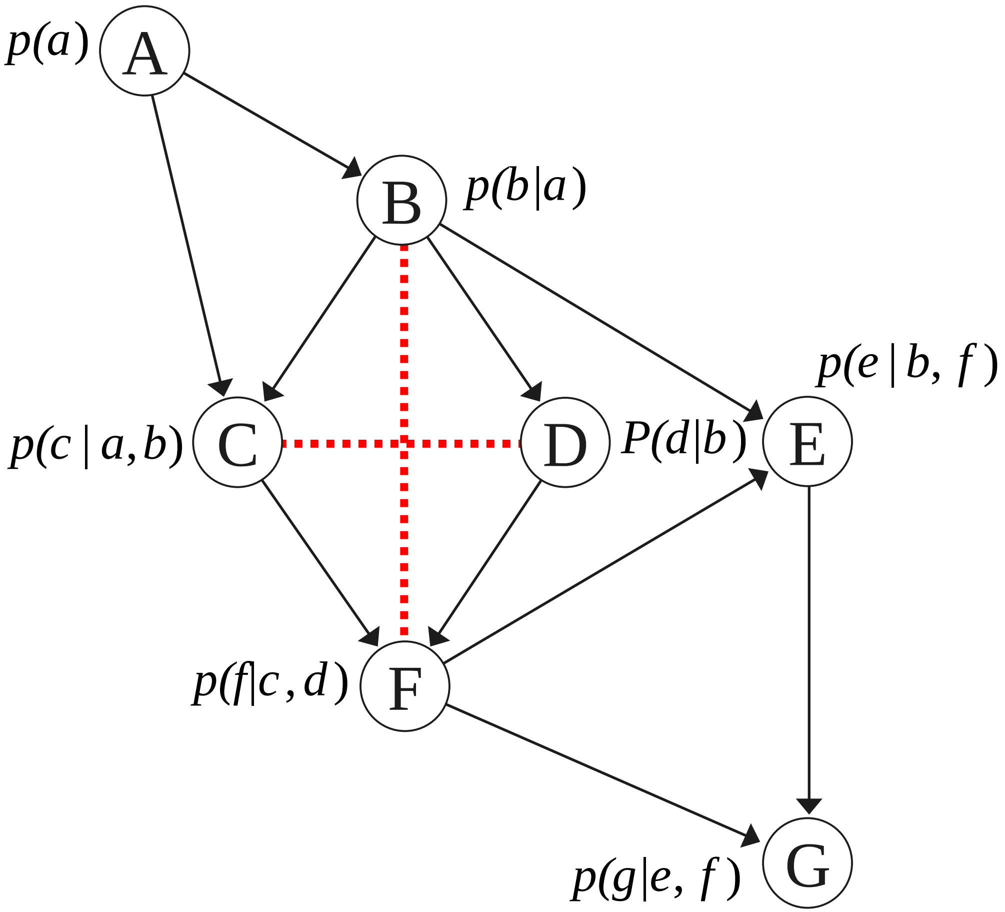
- Correctness and completeness: Algorithm CTE is correct, i.e. it computes the exact joint probability of a single variable and the evidence.
- Time complexity:
  - $O(deg \times (n+N) \times d^{w^*+1})$
- Space complexity: $O(N \times d^{sep})$

where
deg = the maximum degree of a node
n = number of variables (= number of CPTs)
N = number of nodes in the tree decomposition
d = the maximum domain size of a variable
*w** = the induced width, treewidth
sep = the separator size

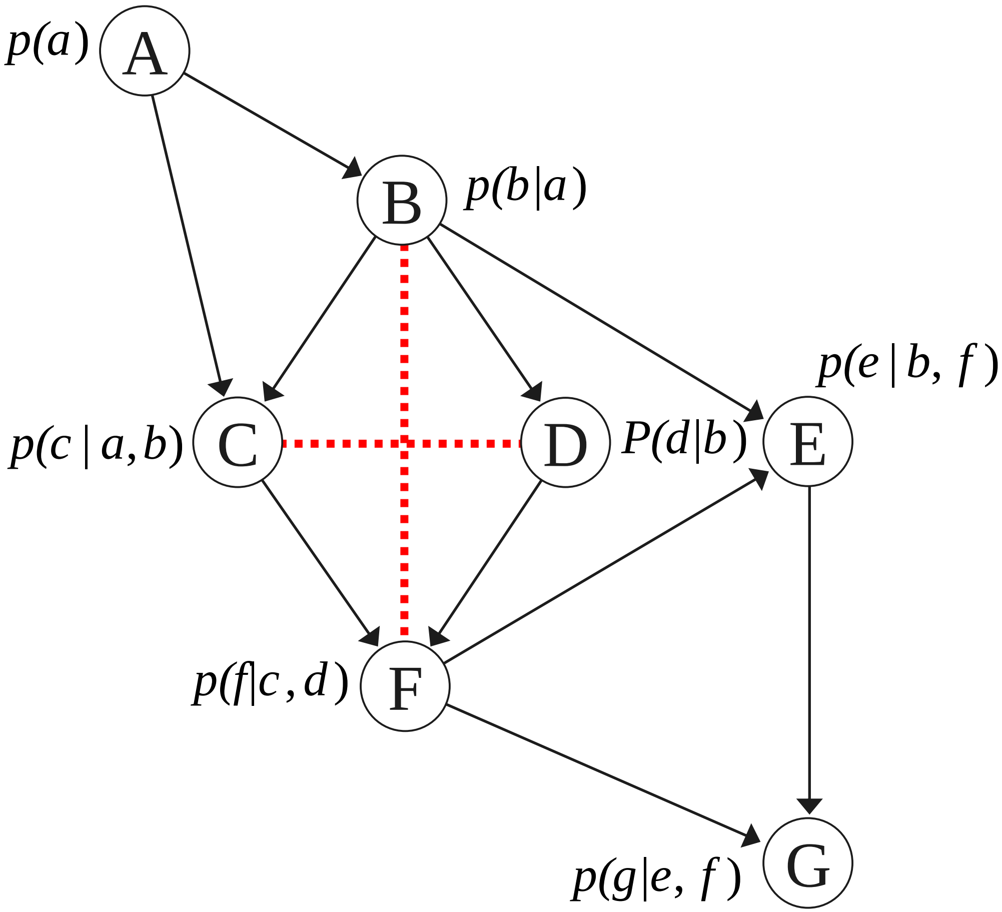
Agenda

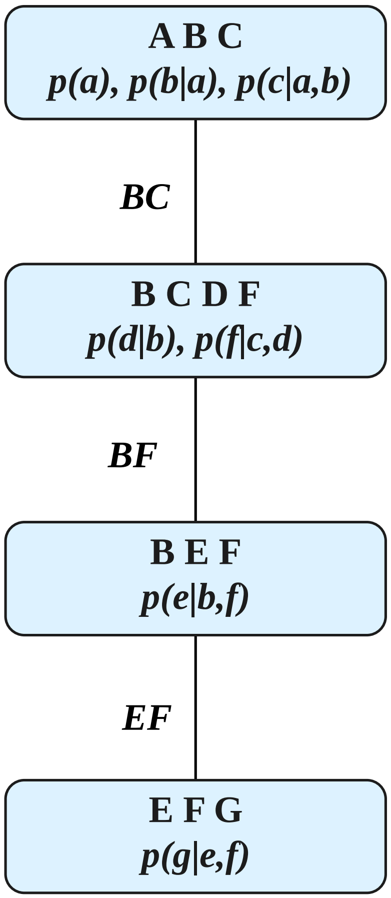
- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

Tree Decomposition for Belief Updating

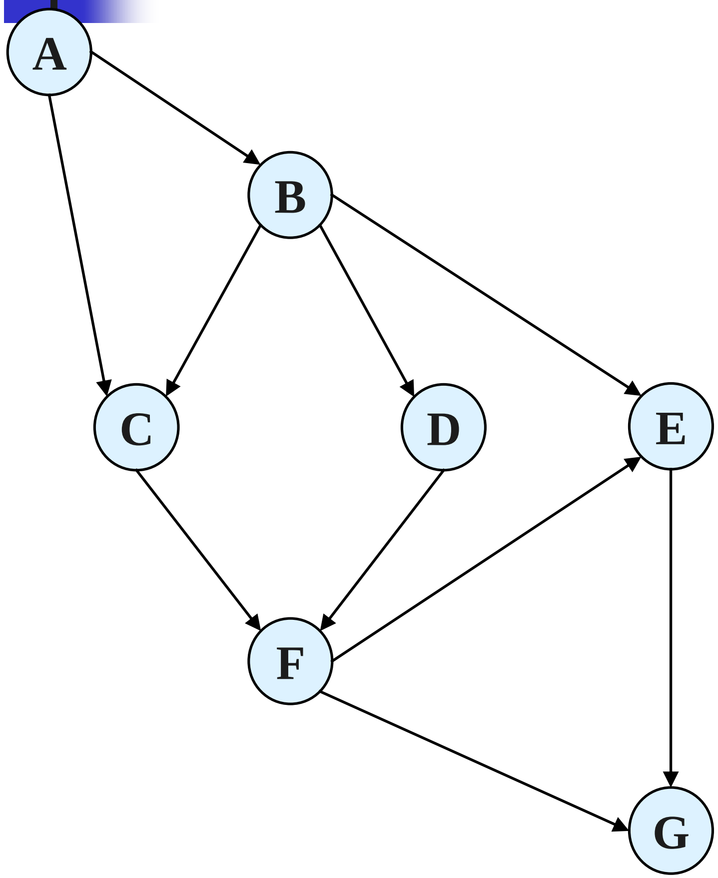


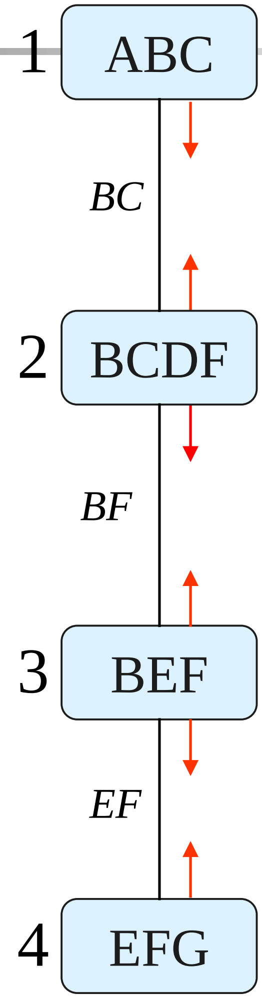
Tree Decomposition for belief updating





CTE: Cluster Tree Elimination





$$h_{(1,2)}(b,c) = \sum_{a} p(a) \cdot p(b \mid a) \cdot p(c \mid a,b)$$

$$h_{(2,1)}(b,c) = \sum_{d,f} p(d \mid b) \cdot p(f \mid c,d) \cdot h_{(3,2)}(b,f)$$

$$h_{(2,3)}(b,f) = \sum_{c,d} p(d \mid b) \cdot p(f \mid c,d) \cdot h_{(1,2)}(b,c)$$

$$h_{(3,2)}(b,f) = \sum_{e} p(e \mid b,f) \cdot h_{(4,3)}(e,f)$$

$$h_{(3,4)}(e,f) = \sum_{b} p(e \mid b,f) \cdot h_{(2,3)}(b,f)$$

$$h_{(4,3)}(e,f) = p(G = g_e \mid e,f)$$

Time: *O (exp(w+1))*
Space: *O (exp(sep))*

For each cluster P(X|e) is computed, also P(e)

Example

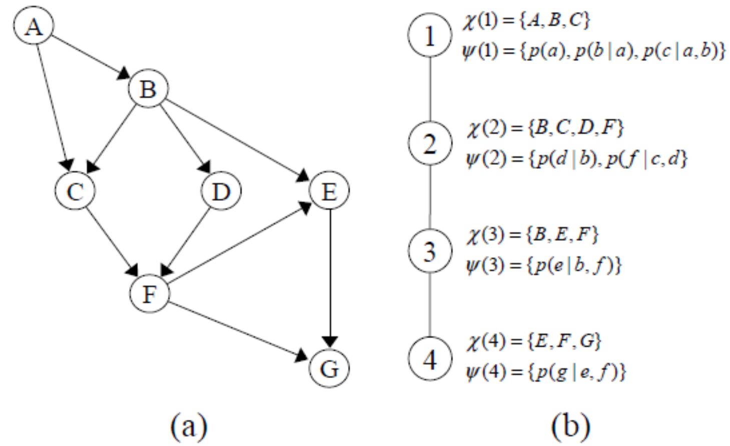


Figure 6.12: [Execution of CTE-BU]: *a*) A belief network; *b*) A join-tree decomposition; *c*)Execution of CTE-BU; no individual functions appear in this case (d) the explicit tree-decomposition

Example

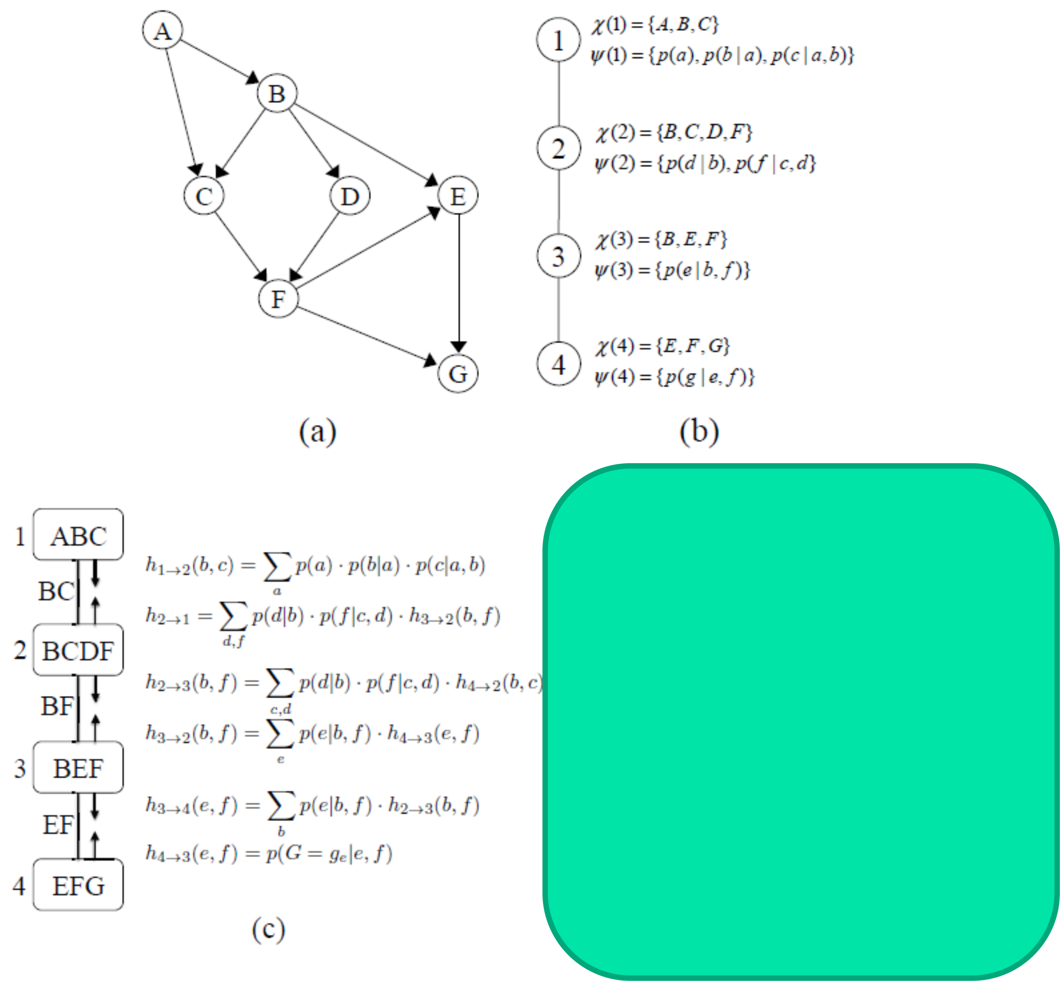


Figure 6.12: [Execution of CTE-BU]: *a*) A belief network; *b*) A join-tree decomposition; *c*)Execution of CTE-BU; no individual functions appear in this case (d) the explicit tree-decomposition

Example

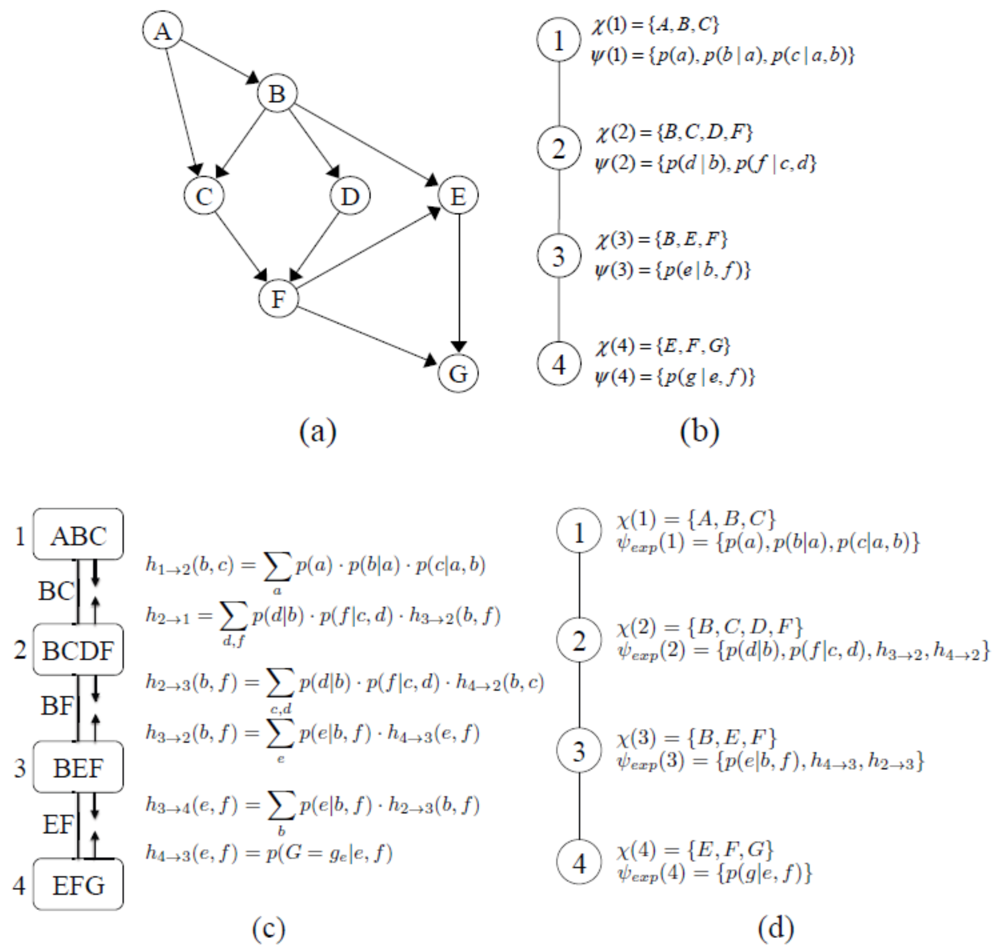


Figure 6.12: [Execution of CTE-BU]: *a*) A belief network; *b*) A join-tree decomposition; *c*)Execution of CTE-BU; no individual functions appear in this case (d) the explicit tree-decomposition

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

Polytrees and Acyclic Networks

- **Polytree:** a BN whose undirected skeleton is a tree
- **Acyclic network**: A network is acyclic if it has a tree-decomposition where each node has a single original CPT.
- A polytree is an acyclic model.

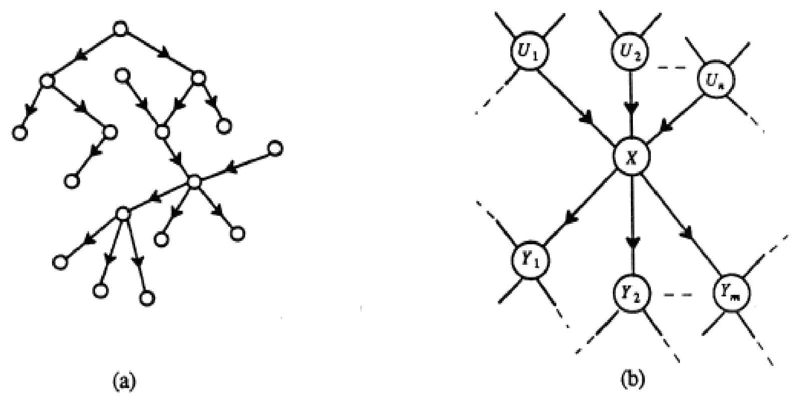
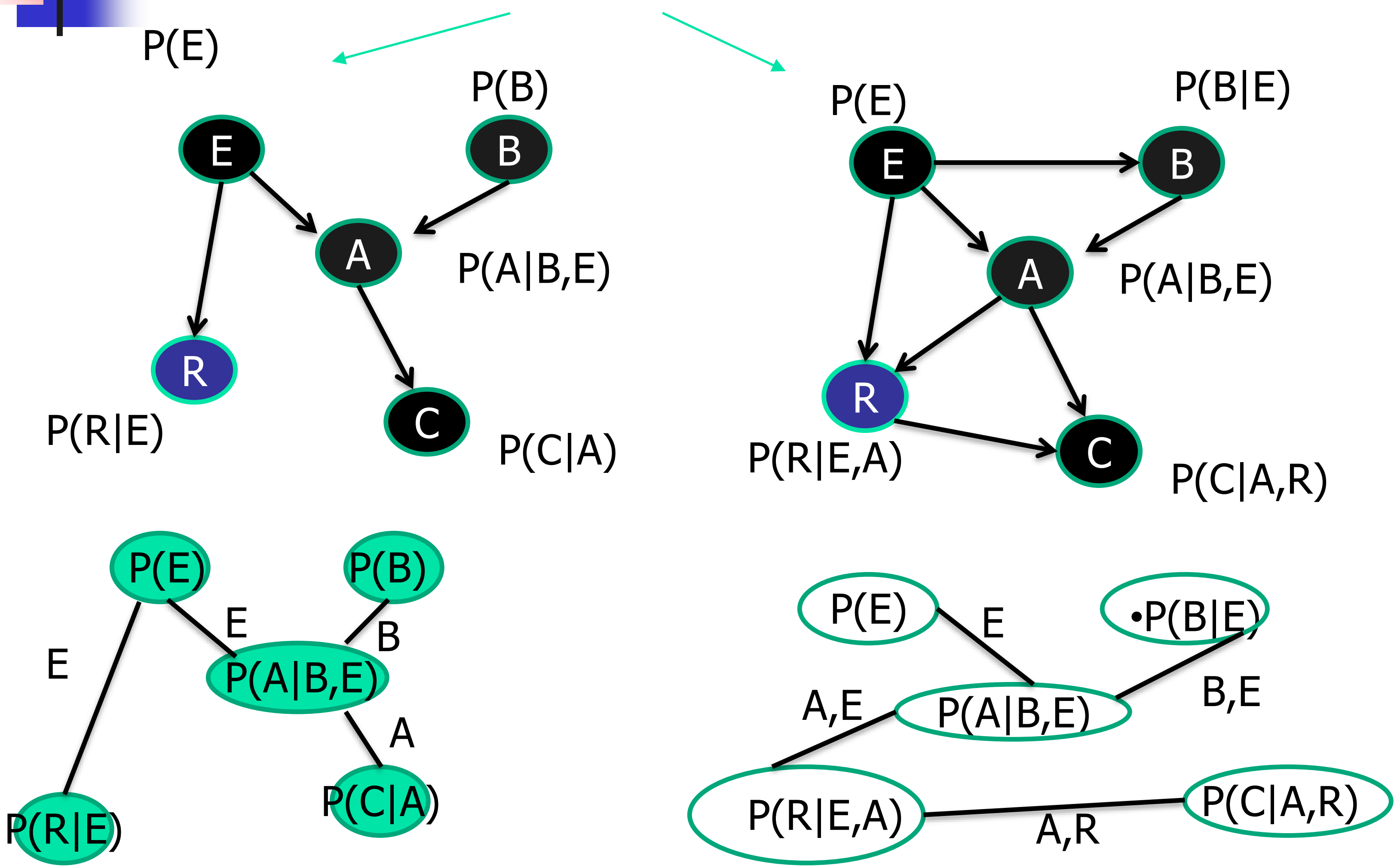


Figure 4.18. (a) A fragment of a polytree and (b) the parents and children of a typical node X.

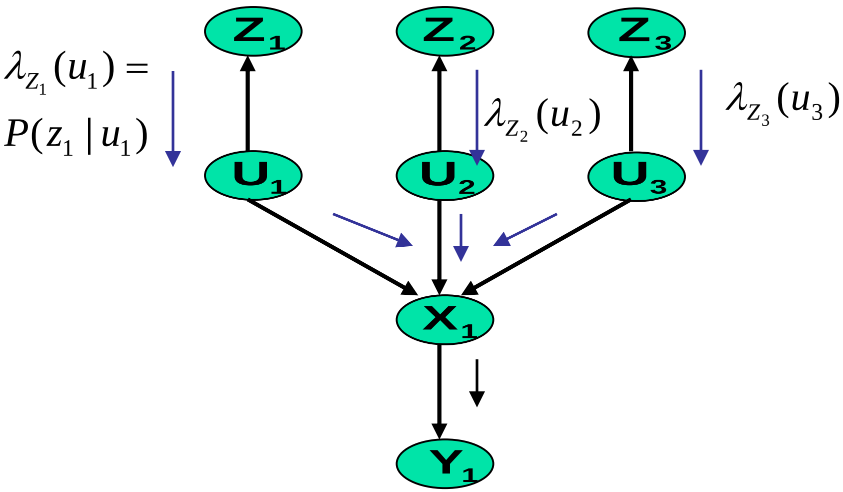
PolyTrees and Acyclic-Networks:

Two acyclic Bayesian Networks

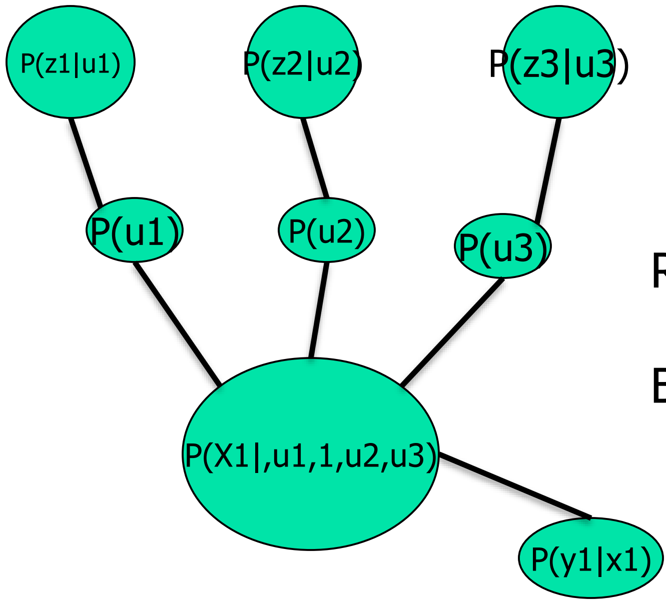


BTE = Belief Propagation is Easy on Polytrees: Pearl's Belief Propagation

A polytree: a tree with
Larger families



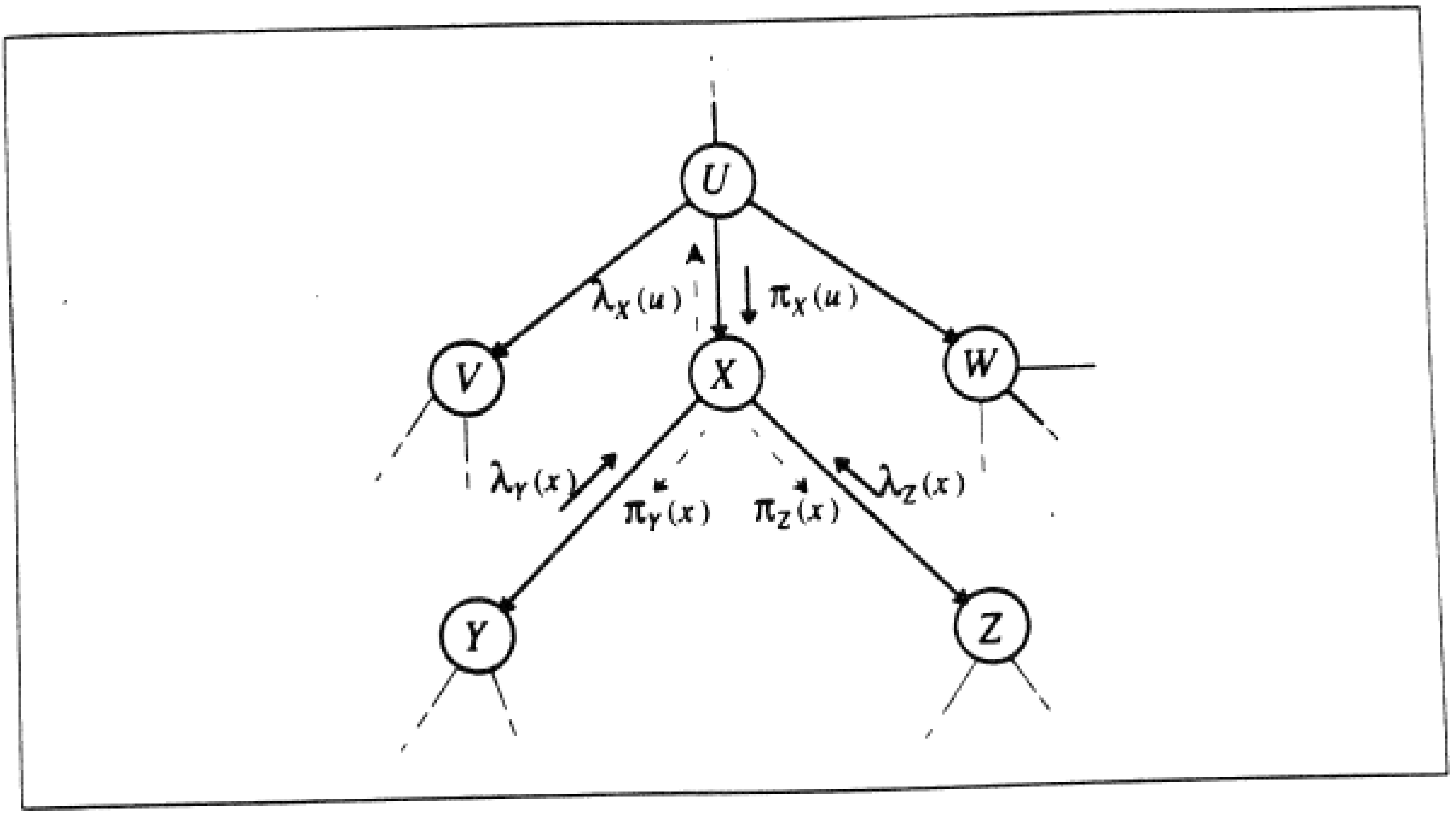
A polytree decomposition



Running BTE = running Pearl's BP over the dual graph

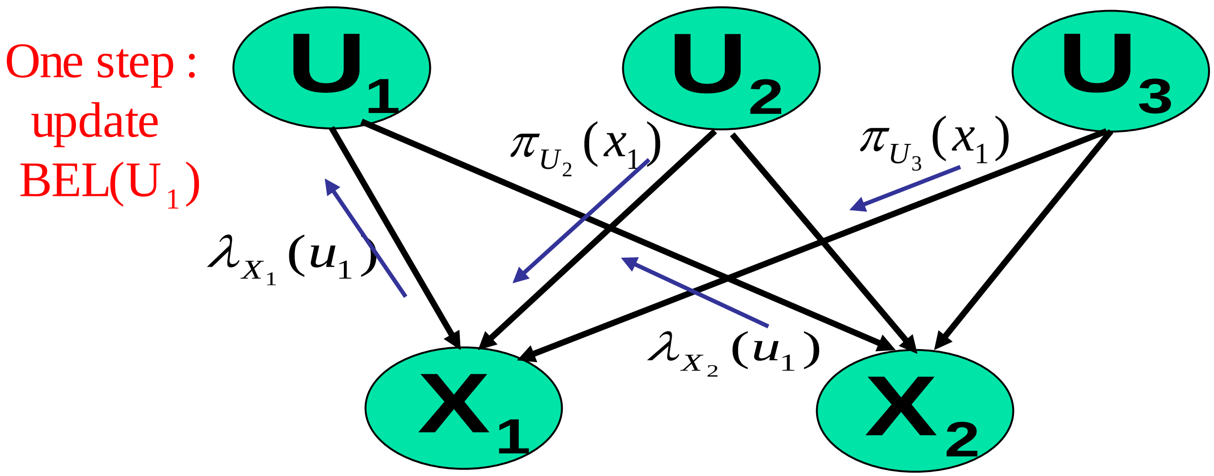
BP is Time and space linear

Pearl's Belief Propagation



From Exact to Approximate: Iterative Belief Propagation

- Belief propagation is exact for poly-trees
- IBP - applying BP iteratively to cyclic networks



- No guarantees for convergence
- Works well for many coding networks

Iterative Belief Propagation

Algorithm IBP

Input: An arc-labeled dual join-graph $DJ = (V, E, L)$ for a graphical model $\mathcal{M} =< X, D, F, \prod >$.

Output: An augmented graph whose nodes include the original functions and the messages received from neighbors. Denote by: h_u^v the message from u to v; $ne(u)$ the neighbors of u in V; $ne_v(u) = ne(u) - \{v\}$; l_{uv} the label of $(u,v) \in E$; $elim(u,v) = scope(u) - scope(v)$.

- **One iteration of IBP**

  For every node u in DJ in a topological order and back, do:

  1. **Process observed variables**

     Assign evidence variables to the each p_i and remove them from the labeled arcs.

  2. **Compute and send to v the function:**

$$h_u^v = \sum_{elim(u,v)} (p_u \cdot \prod_{\{h_i^u, i \in ne_v(u)\}} h_i^u)$$

  Endfor

- **Compute approximations of $P(F_i|e)$, $P(X_i|e)$:**

  For every $X_i \in X$ let u be the vertex of family F_i in DJ,

  $P(F_i|e) = \alpha(\prod_{h_i^u, u \in ne(i)} h_i^u) \cdot p_u$;

  $P(X_i|e) = \alpha \sum_{scope(u)-\{X_i\}} P(F_i|e)$.

Agenda

- From bucket-elimination (BE) to bucket-tree elimination (BTE)
- From BTE to CTE, Acyclic networks, the join-tree algorithm
- Generating join-trees, the treewidth
- Examples of CTE for Bayesian network
- Belief-propagation on acyclic probabilistic networks (poly-trees) and Loopy networks
- Conditioning with elimination (Dechter, 7.1, 7.2)

The idea of cutset-conditioning

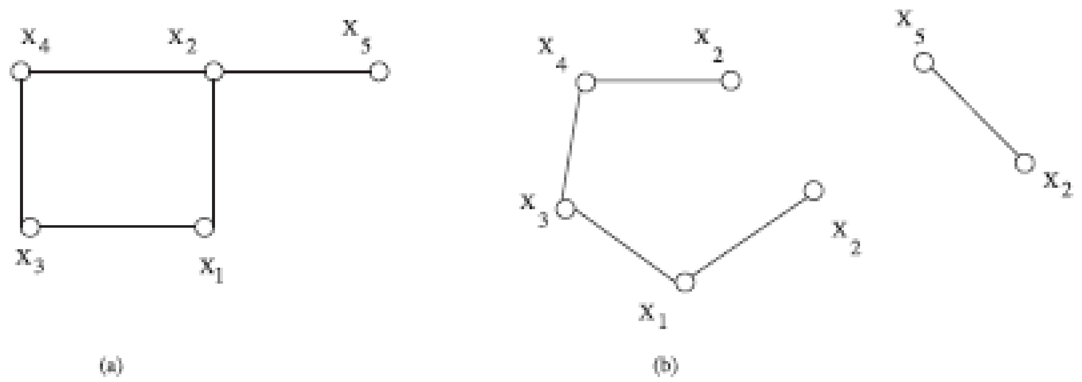
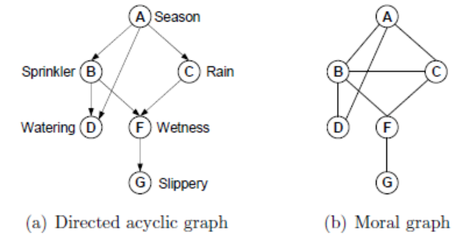


Figure 7.1: An instantiated variable cuts its own cycles.

The impact of observations



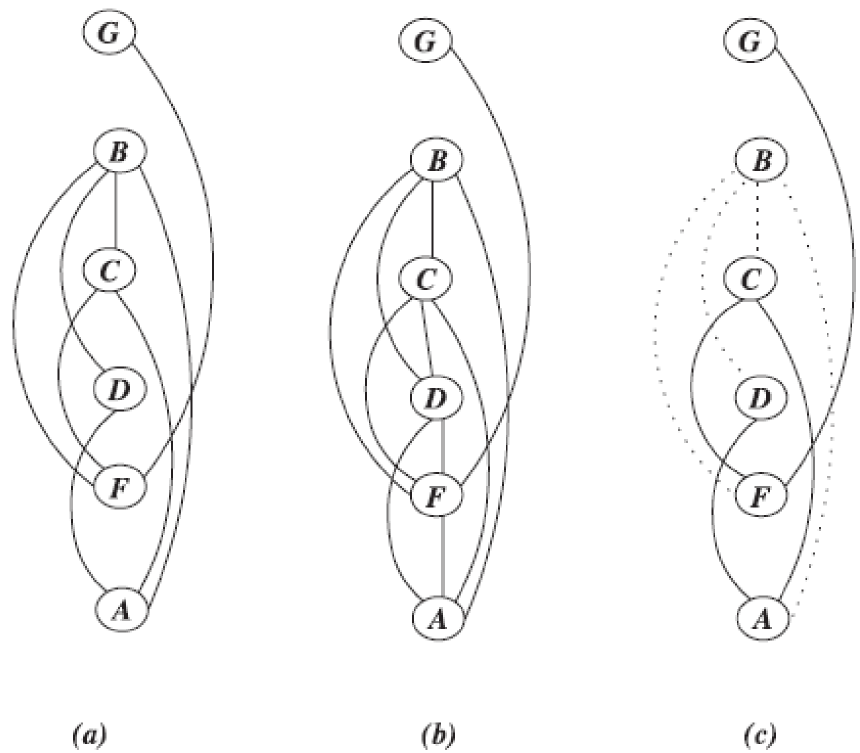


Figure 4.9: Adjusted induced graph relative to observing B.

Ordered graph   Induced graph   Ordered conditioned graph

The Idea of Cutset-Conditioning

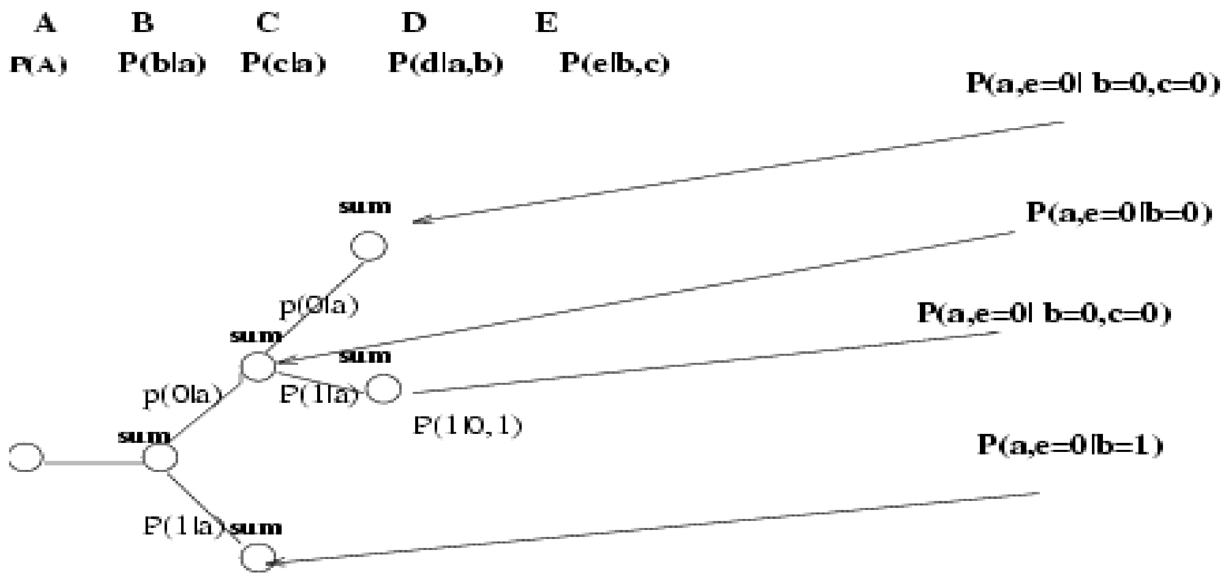
We observed that when variables are assigned, connectivity reduces.
The magnitude of saving is reflected through the "conditioned-induced graph"

- Cutset-conditioning exploit this in a systematic way:
- Select a subset of variables, assign them values, and
- Solve the conditioned problem by BE.
- Repeat for all assignments to the cutset.

Algorithm VEC

Conditioning+Elimination

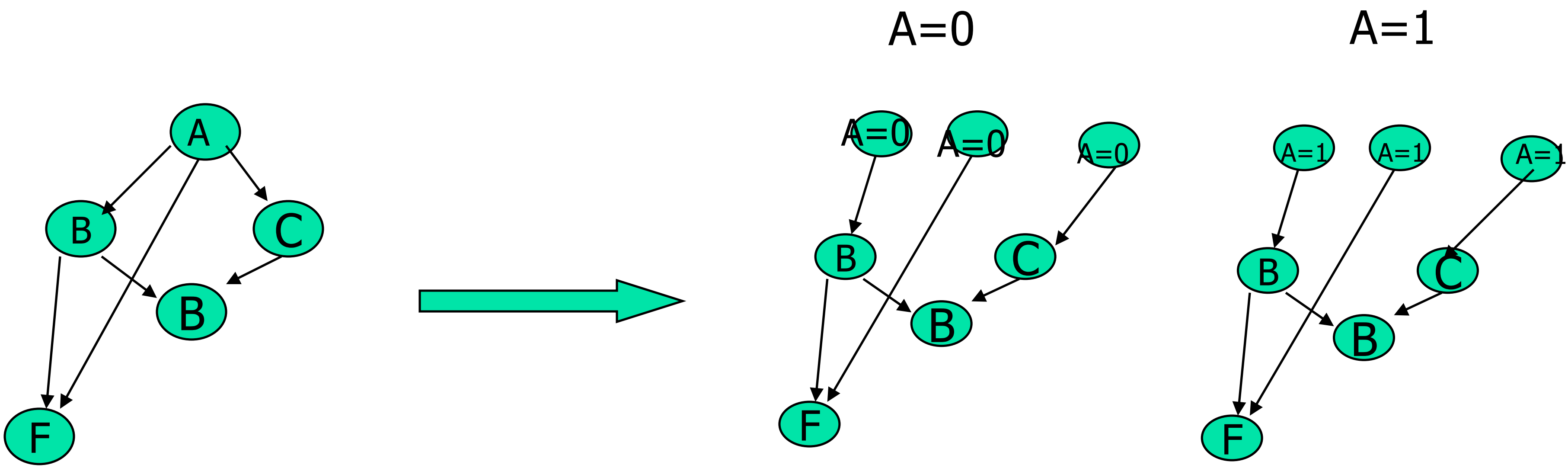
$$P(a,e=0) = P(a)\sum_b P(b|a)\sum_c P(c|a)\sum_d P(d|a,b)\sum_{e=0} P(e|b,c)$$



Idea: conditioning until w^* of a (sub)problem gets small

Loop-Cutset Conditioning

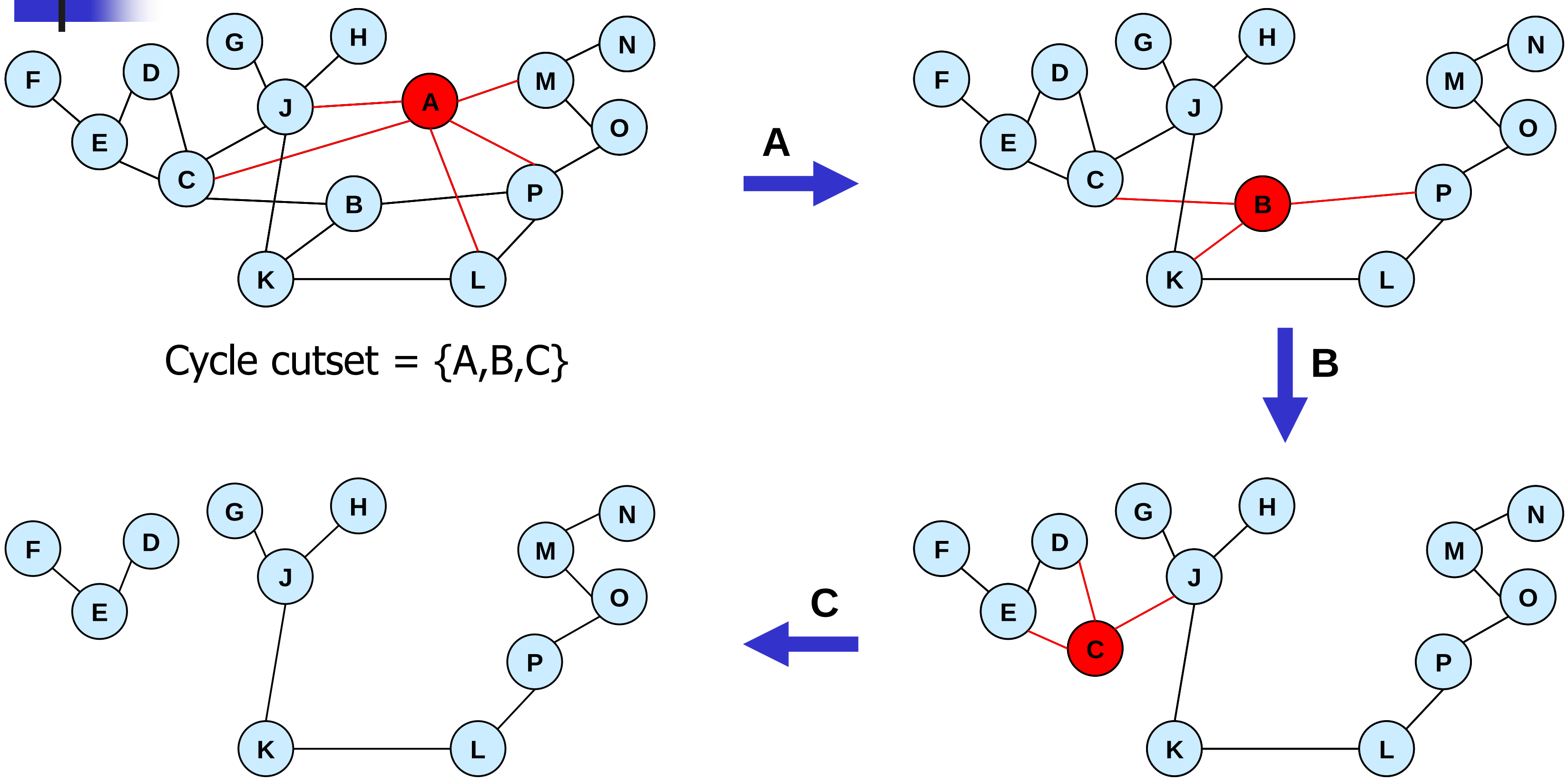
- You condition until you get a polytree



P(B|F=0) = P(B, A=0|F=0)+P(B,A=1|F=0)

Loop-cutset method is time exp in loop-cutset size and linear space. For each cutset we can do BP

Cycle-Cutset Conditioning



1-cutset = {A,B,C}, size 3

q-Cutset, Minimal

Definition 7.3 ***q*-cutset, minimal.** Given a graph G, a subset of nodes is called a q-*cutset* for an integer q iff when removed, the resulting graph has an induced-width less than or equal to q. A minimal q-*cutset* of a graph has a smallest size among all q-cutsets of the graph. A cycle-cutset is a 1-cutset of a graph.

Finding a minimal q-cutset is clearly a hard task [A. Becker and Geiger, 1999; Bar-Yehuda *et al.*, 1998; Becker *et al.*, 2000; Bidyuk and Dechter, 2004]. However, like in the special case of a cycle-cutset we can settle for a non-minimal q-cutset relative to a given variable ordering. Namely,

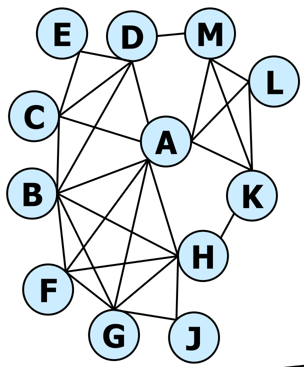
Example 7.4 Consider as another example the contsaint graph of a graph coloring problem given in Figure 7.3a. The search space over a 2-cutset, and the induced-graph of the conditioned instances are depicted in 7.3b.

Loop-Cutset, q-Cutset, cycle-cutset

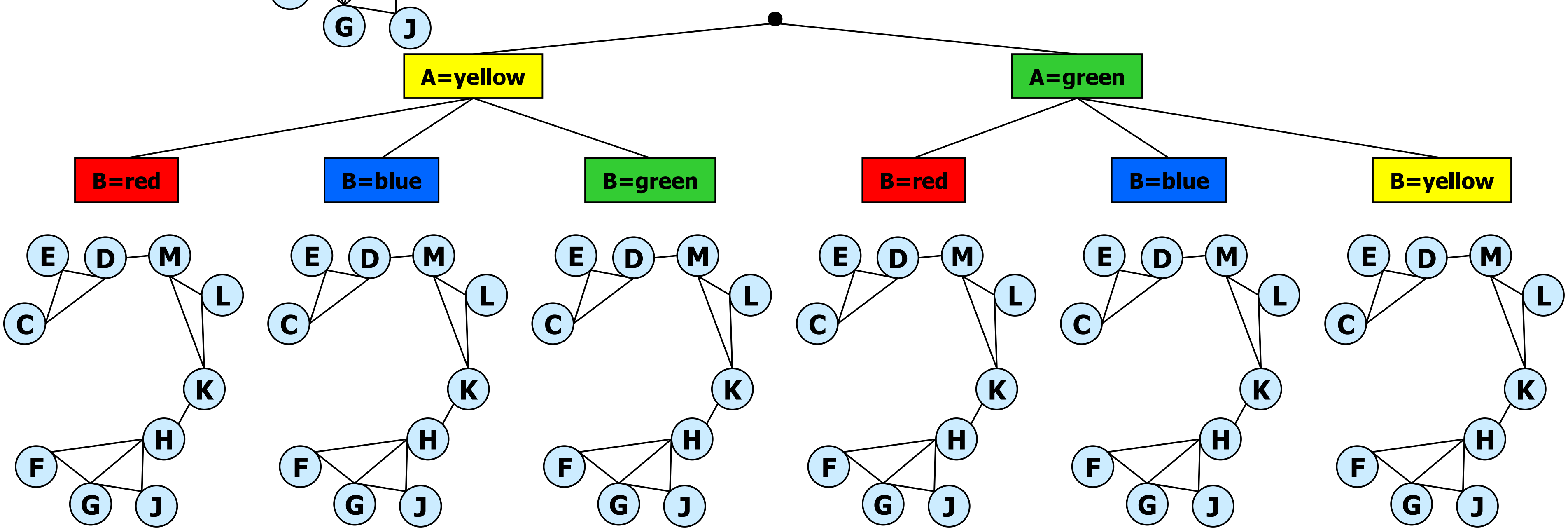
- A loop-cutset is a subset of nodes of a directed graph that when removed the remaining graph is a poly-tree
- A q-cutset is a subset of nodes of an undirected graph that when removed the remaining graph is has an induced-width of q or less.
- A cycle-cutset is a q-cutset such that q=1.

Search Over the Cutset (cont)

Graph Coloring problem



- Inference may require too much memory
- **Condition** on some of the variables



2-cutset = {A,B}, size =2

VEC: Variable Elimination with Conditioning; or, q-cutset lgorithms

- VEC-bel:
- Identify a q-cutset, C, of the network
- For each assignment to C=c solve by CTE or BE the conditioned sub-problem.
- Accumulate probability.
- Time complexity: nk^{c+q+1}
- Space complexity: nk^{q}

Algorithm VEC (Variable-elimination with conditioning)

ALGORITHM VEC-EVIDENCE

Input: A belief network $\mathcal{B} =< \mathcal{X}, \mathcal{D}, \mathcal{G}, \mathcal{P} >$, an ordering $d = (x_1, \ldots, x_n)$; evidence e over E, a subset C of conditioned variables;

output: The probability of evidence $P(e)$

Initialize: $\lambda = 0$.

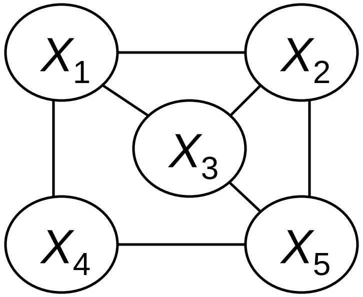
1. For every assignment $C = c$, do
   - $\lambda_1 \leftarrow$ The output of BE-bel with $c \cup e$ as observations.
   - $\lambda \leftarrow \lambda + \lambda_1$. (update the sum).

2. **Return** $P(e) = \alpha \cdot \lambda$ (α is a normalization constant.)

VEC and ALT-VEC: Alternate conditioning and Elimination

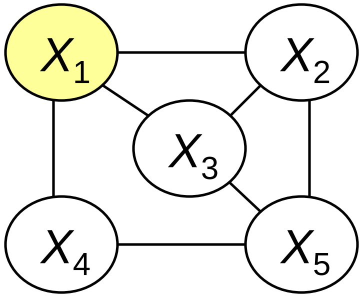
- VEC (q-cutset-conditioining) can also alternate search and elimination, yielding ALT-VEC.
- A time-space tradeoff

Search Basic Step: Conditioning

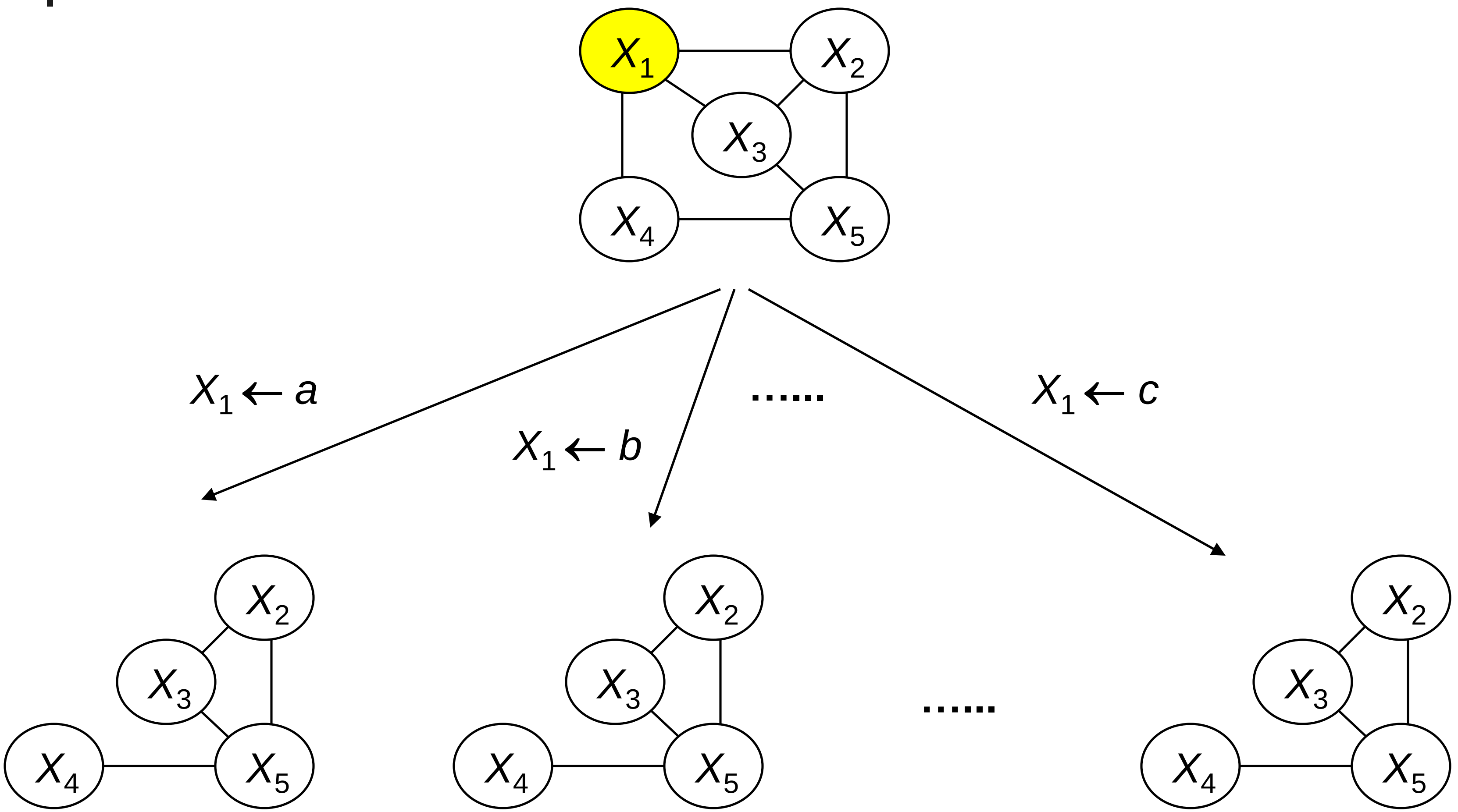


Search Basic Step: Conditioning

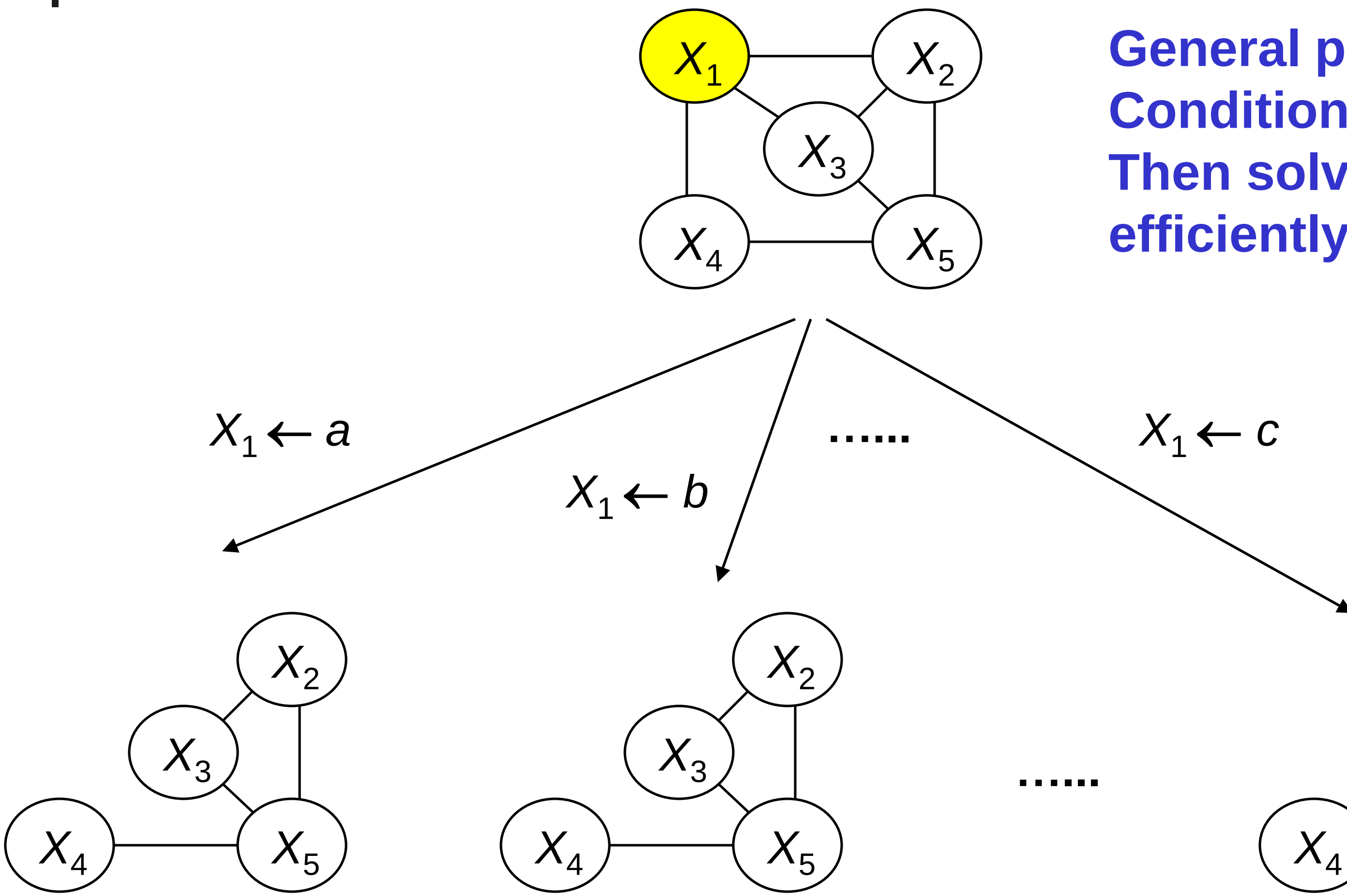
- Select a variable



Search Basic Step: Conditioning

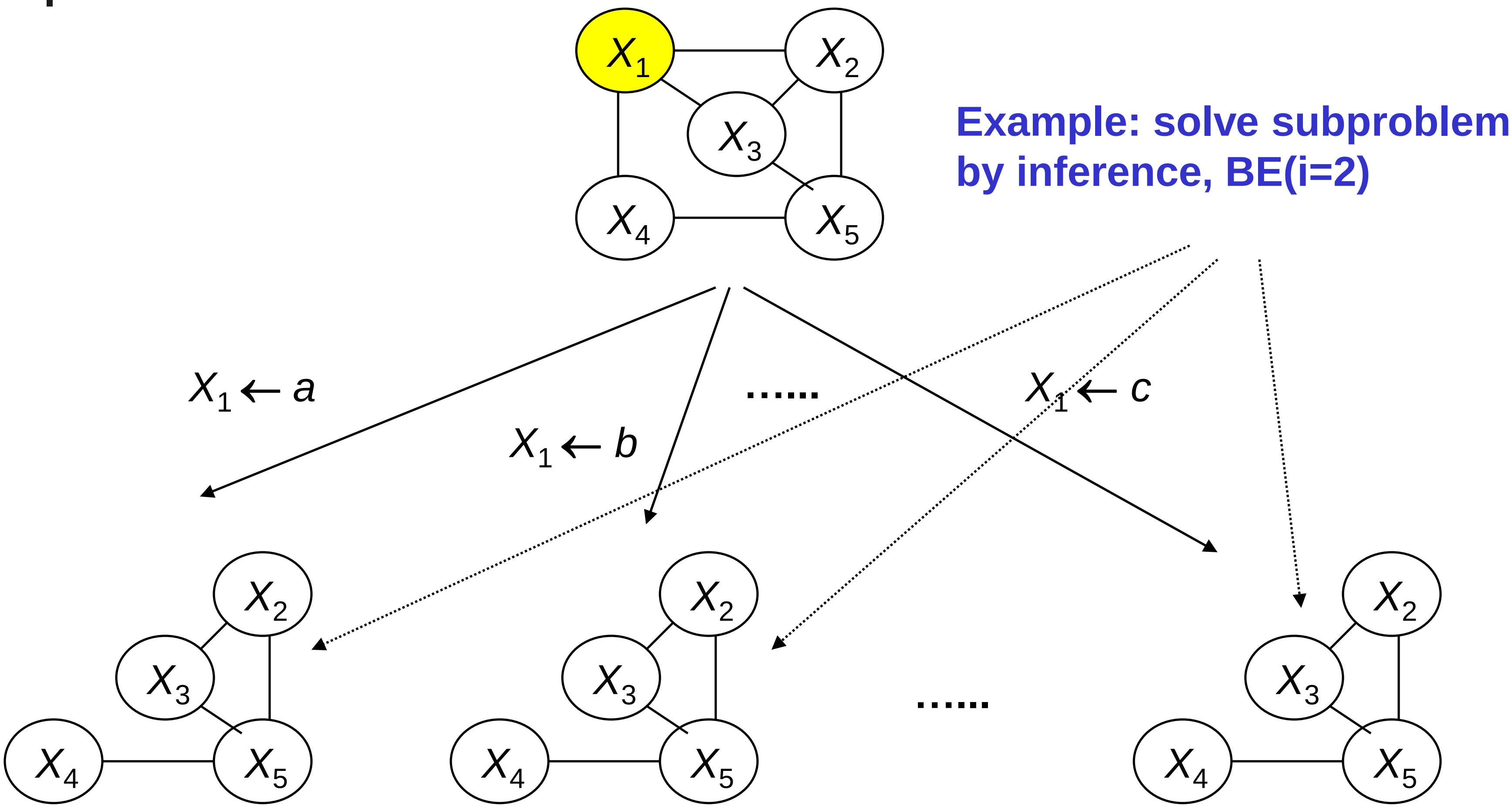


Search Basic Step:
Variable Branching by Conditioning



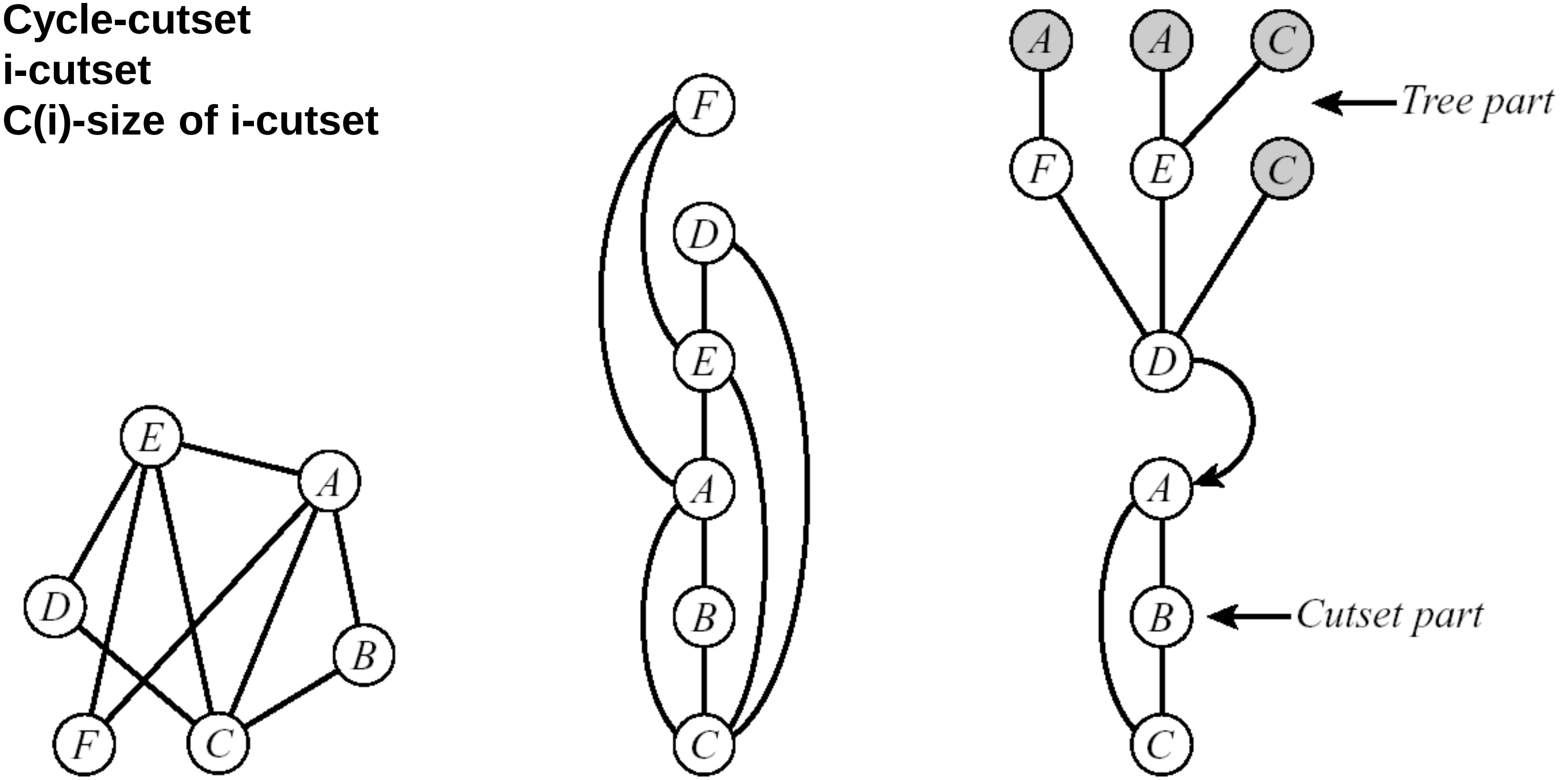
General principle:
Condition until tractable
Then solve sub-problems efficiently

Search Basic Step:
Variable Branching by Conditioning



The Cycle-Cutset Scheme:
Condition Until Treeness

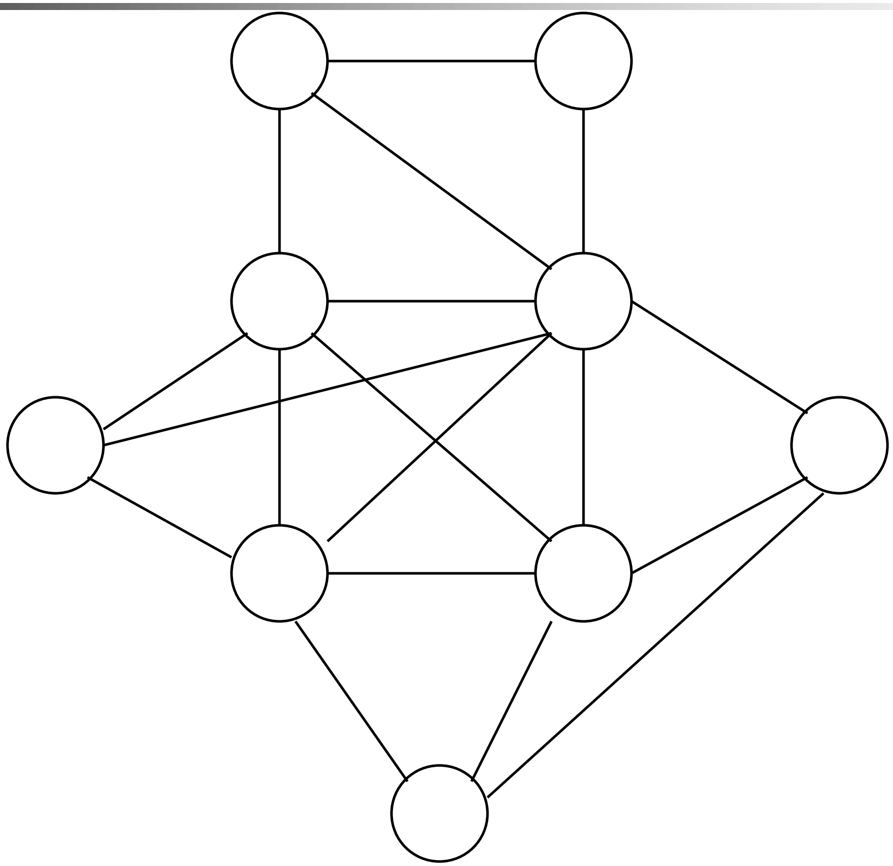
- **Cycle-cutset**
- **i-cutset**
- **C(i)-size of i-cutset**



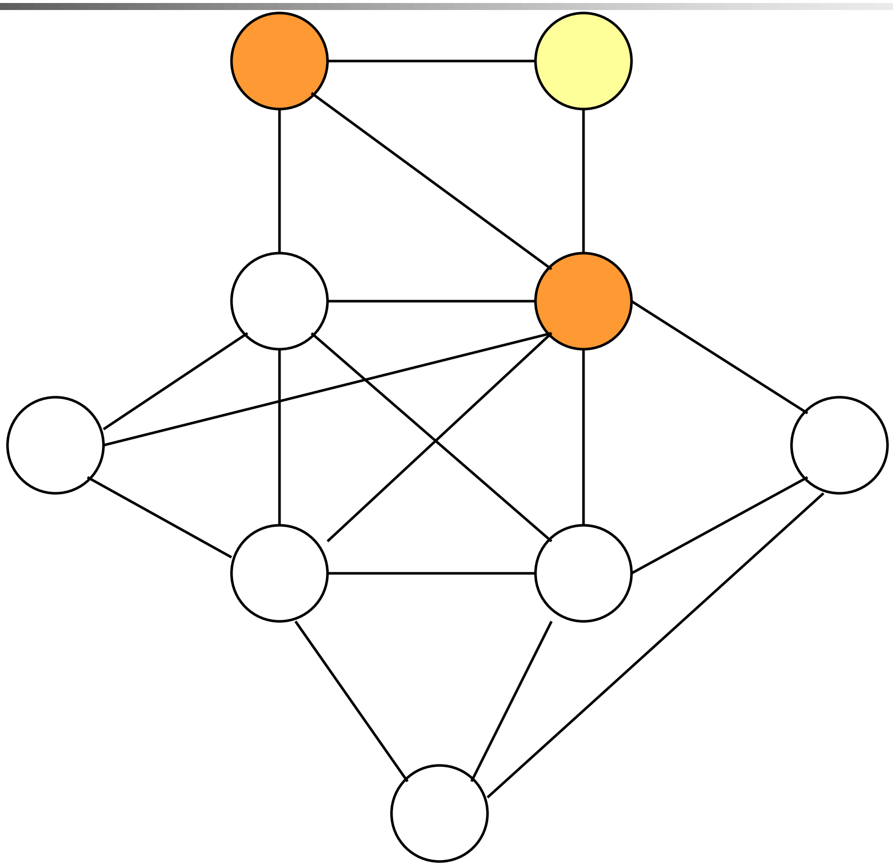
(a) (b) (c)

Space: exp(i), Time: O(exp(i+c(i))

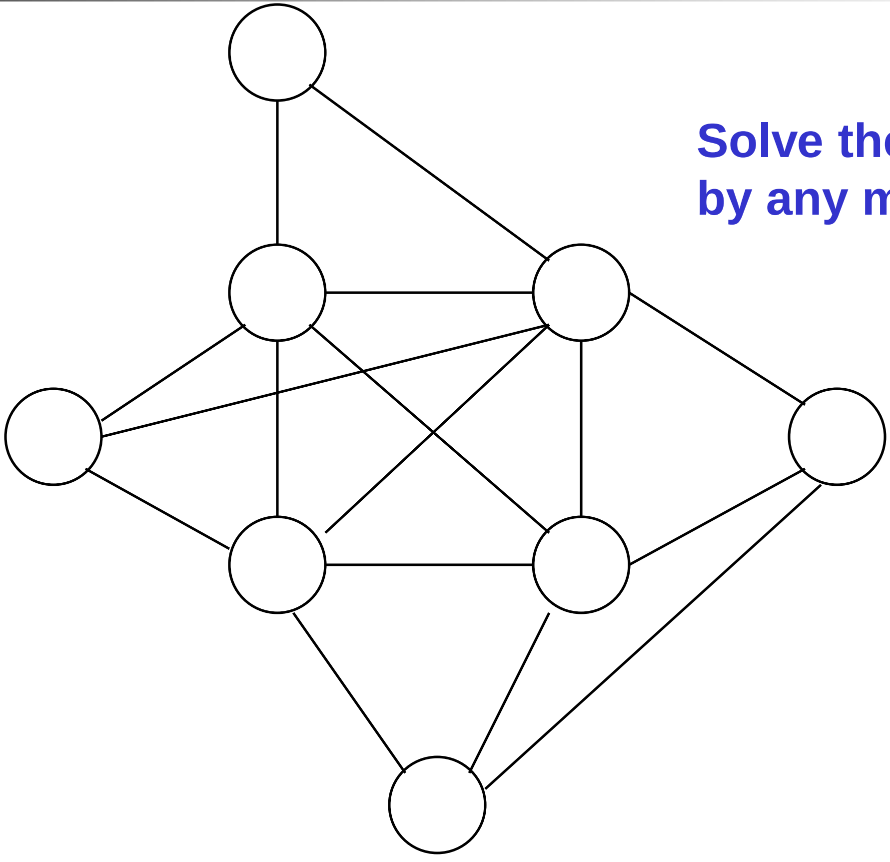
Eliminate First



Eliminate First



Eliminate First

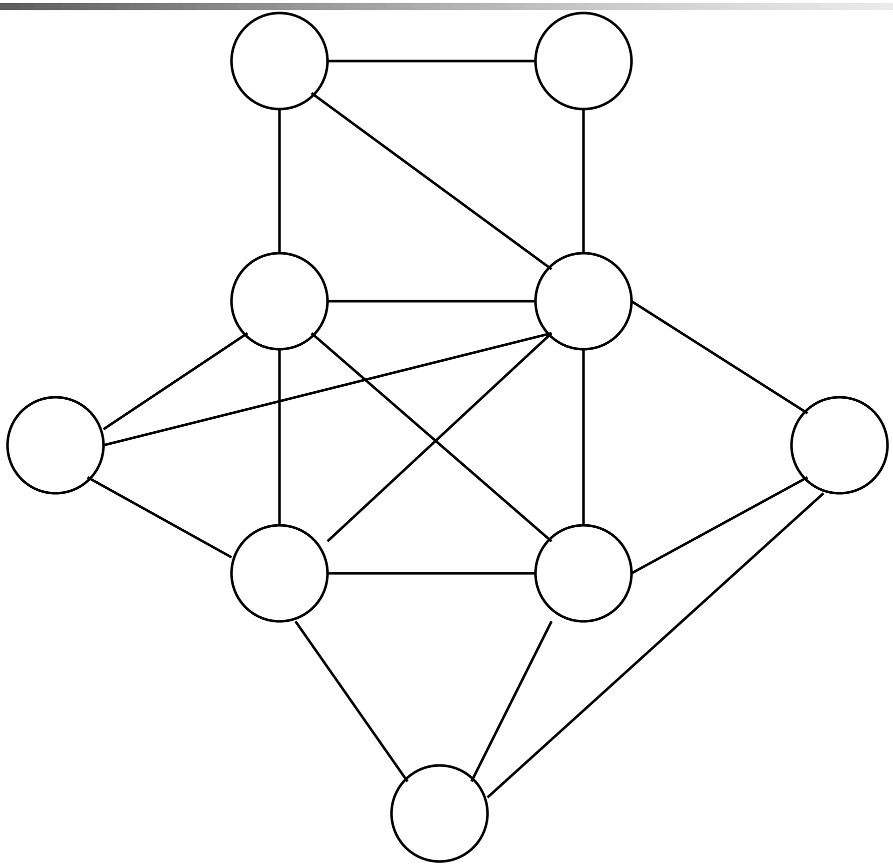


Solve the rest of the problem by any means

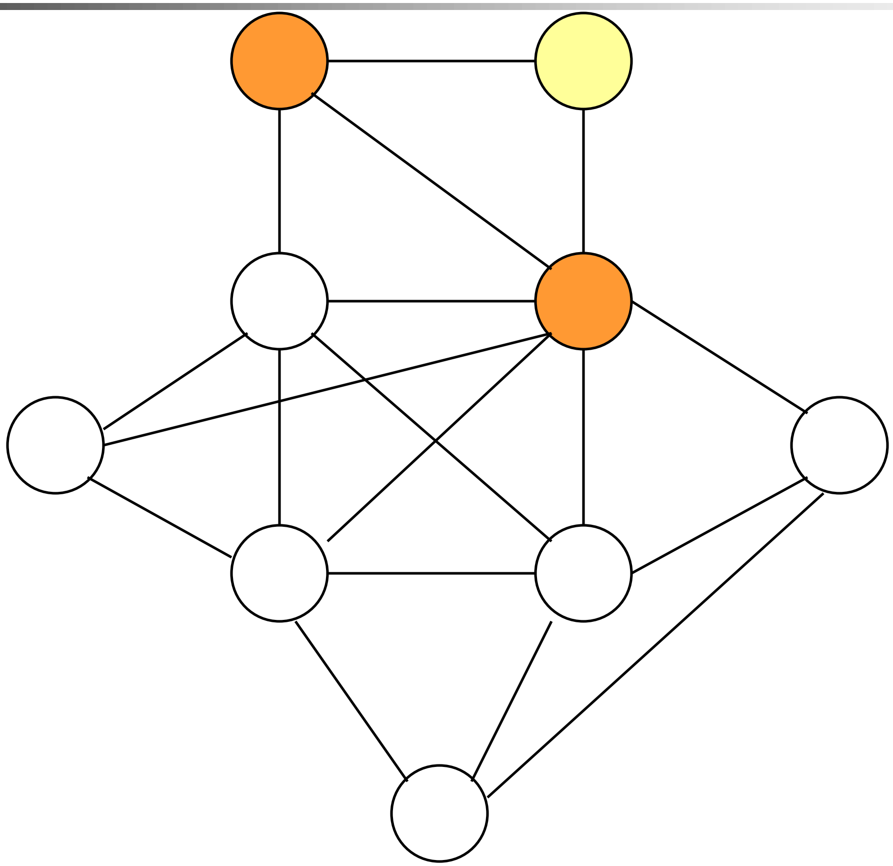
Hybrids Variants

- Condition, condition, condition … and then only eliminate (w-cutset, cycle-cutset)
- Eliminate, eliminate, eliminate … and then only search
- Interleave conditioning and elimination (elim-cond(i), VE+C)

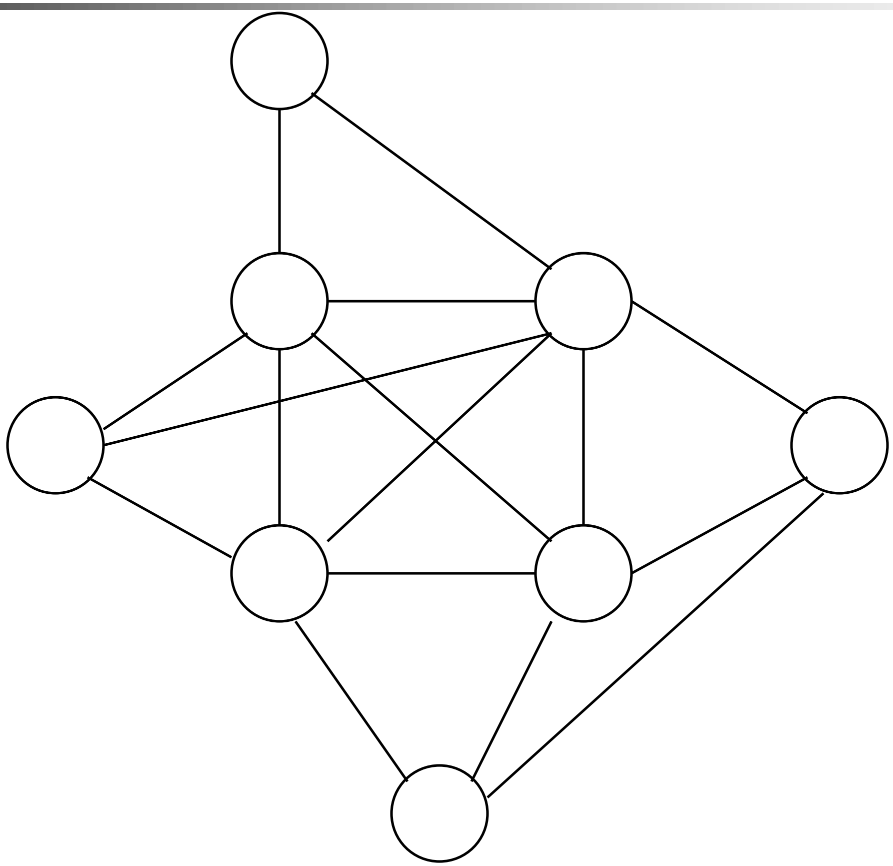
Interleaving Conditioning and Elimination
(Larrosa & Dechter, CP'02)



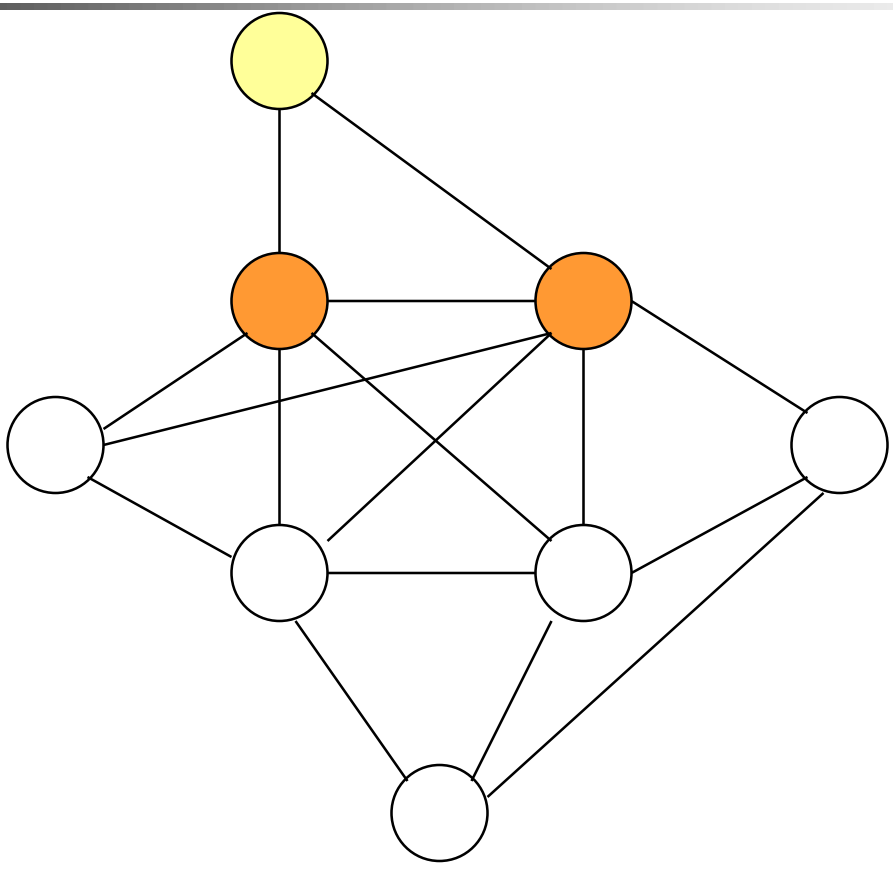
Interleaving Conditioning and Elimination



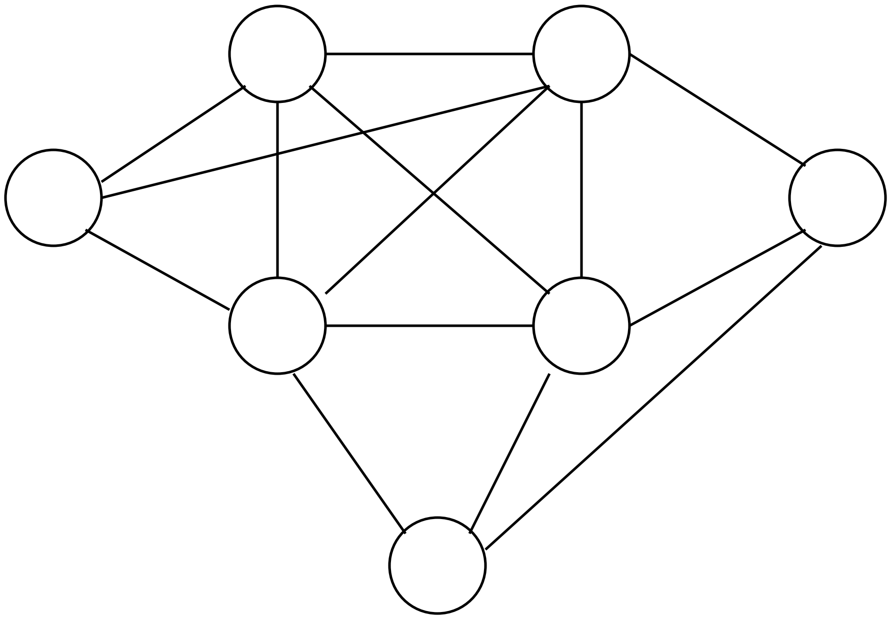
Interleaving Conditioning and Elimination



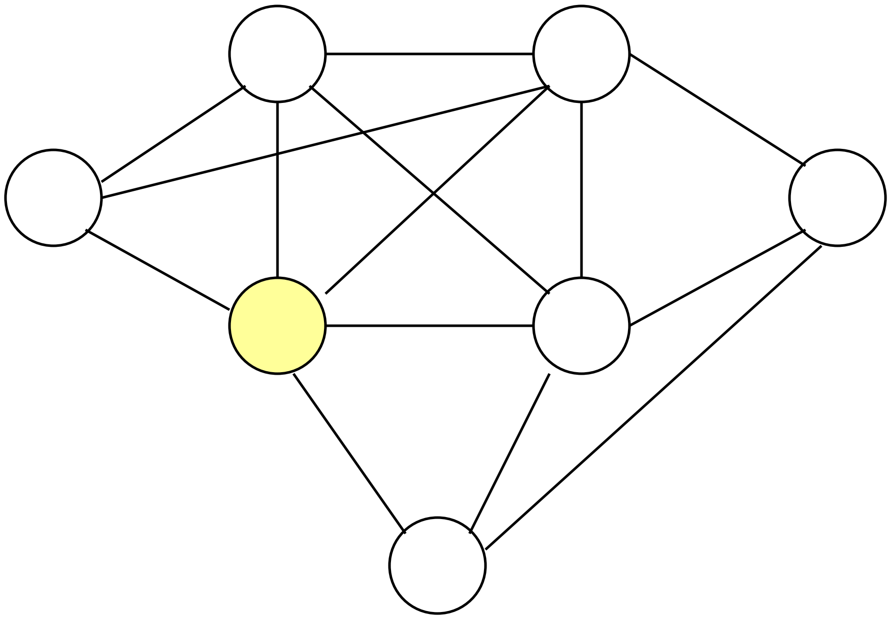
Interleaving Conditioning and Elimination



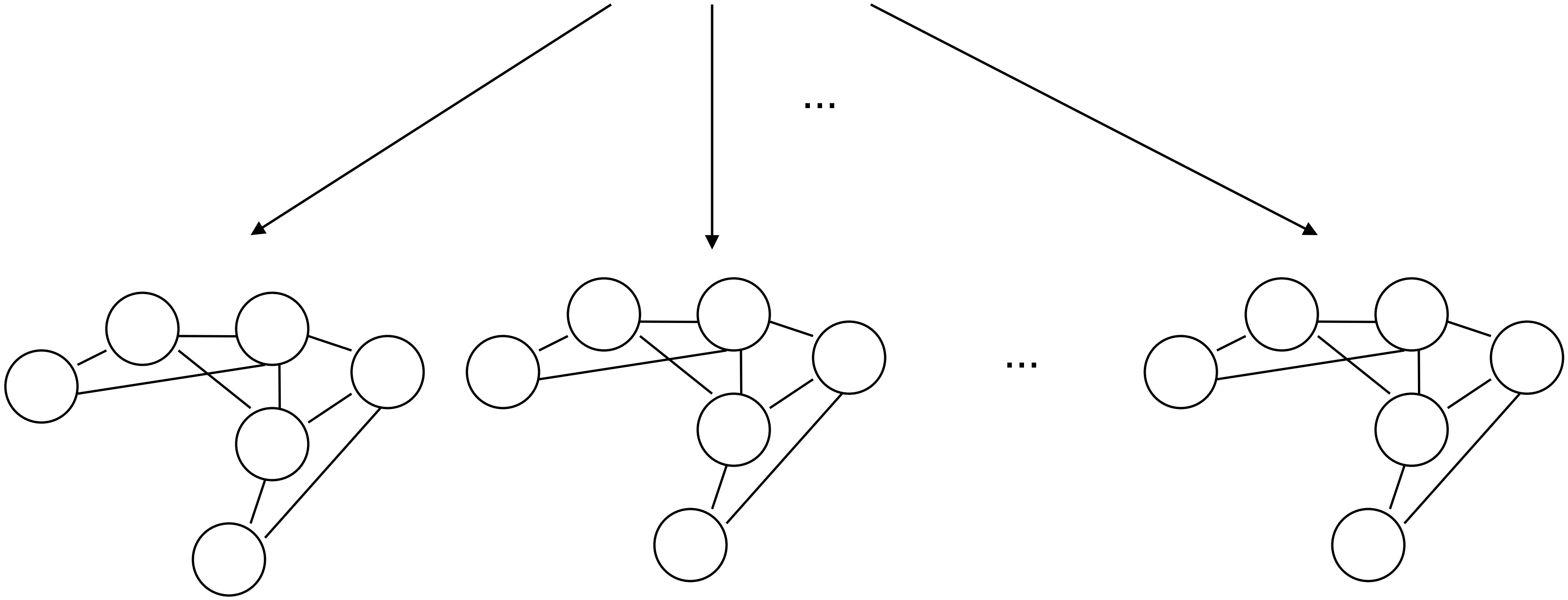
Interleaving Conditioning and Elimination



Interleaving Conditioning and Elimination



Interleaving Conditioning and Elimination



What hybrid should we use?

- q=1? (loop-cutset?)
- q=0? (Full search?)
- $q=w^*$ (Full inference)?
- q in between?
- depends... on the graph
- What is relation between cycle-cutset and the induced-width?

Properties of Conditioning+Elimination

Definition 5.6.1 (cycle-cutset,w-cutset) *Given a graph G, a subset of nodes is called a w-cutset iff when removed from the graph the resulting graph has an induced-width less than or equal to w. A minimal w-cutset of a graph has a smallest size among all w-cutsets of the graph. A cycle-cutset is a 1-cutset of a graph.*

A cycle-cutset is known by the name a *feedback vertex set* and it is known that finding the minimal such set is NP-complete [41]. However, we can always settle for approximations, provided by greedy schemes. Cutset-decomposition schemes call for a new optimization task on graphs:

Definition 5.6.2 (finding a minimal w-cutset) *Given a graph $G = (V, E)$ and a constant w, find a smallest subset of nodes U, such that when removed, the resulting graph has induced-width less than or equal w.*

Tradeoff between w* and q-cutstes

Theorem 7.7 *Given graph G, and denoting by c_q^* its minimal q-cutset then,*

$$1 + c_1^* \geq 2 + c_2^* \geq ...q + c_q^*, ... \geq w^* + c_{w*}^* = w^*.$$

Proof. Let's assume that we have a q-cutset of size c_q. Then if we remove it from the graph the result is a graph having a tree decomposition whose treewidth is bounded by q. Let's T be this decomposition where each cluter has size $q + 1$ or less. If we now take the q-cutset variables and add them back to every cluster of T, we will get a tree decomposition of the whole graph (exercise: show that) whose treewidth is $c_q + q$. Therefore, we showed that for *every* c_q-size q-cutset, there is a tree decomposition whose treewidth is $c_q + q$. In particular, for an optimal q-cutset of size c_q^* we have that $w*$, the treewidth obeys, $w* \leq c_q^* + q$. This does not complete the proof because we only showed that for every q, $w* \leq c_q^* + q$. But, if we remove even a single node from a minimal q-cutset whose size is c_q^*, we get a $q + 1$ cutset by definition, whose size is $c_q^* - 1$. Therefore, $c_{q+1}^* \leq c_q^* - 1$. Adding q to both sides of the last inequality we get that for every $1 \leq q \leq w^*$, $q + c_q^* \geq q + 1 + c_{q+1}^*$, which completes the proof. □