Readings:
K&F: 8.1, 8.2, 8.3, 8.7.1
K&F: 9.1, 9.2, 9.3, 9.4

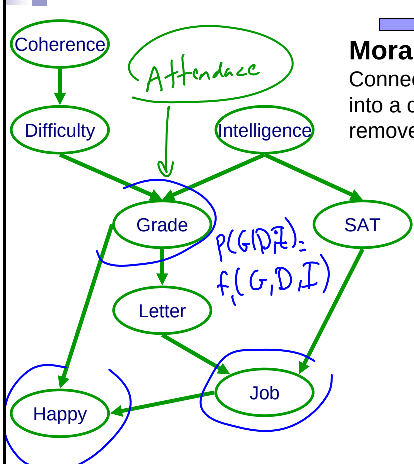
Variable Elimination 2

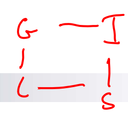
Clique Trees

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

October 13th, 2006

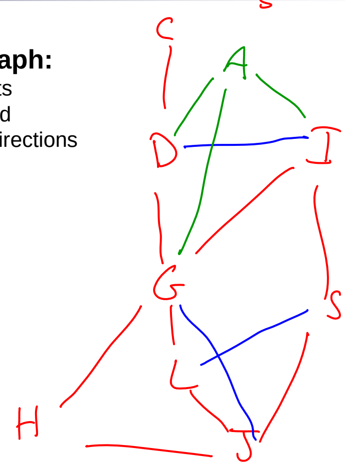
Complexity of variable elimination – Graphs with loops





no chord
=> not chordal graph
(not triangulated)

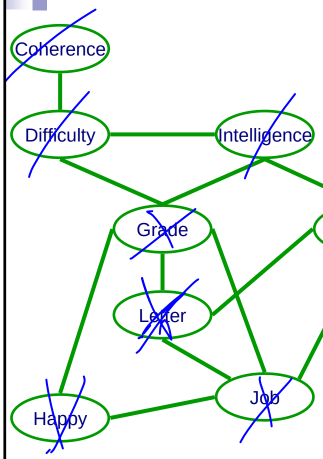
Moralize graph:
Connect parents into a clique and remove edge directions



Connect nodes that appear together in an initial factor

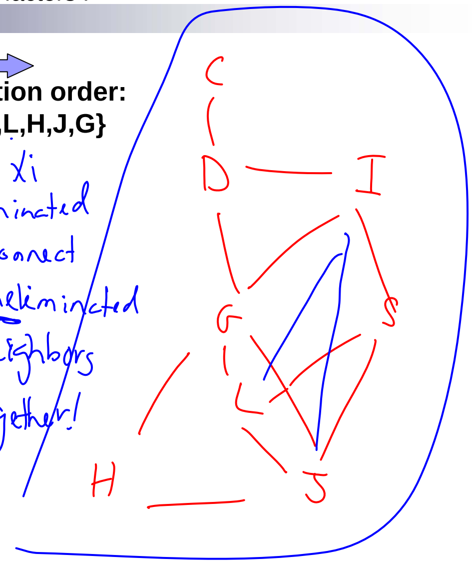
Induced graph

The **induced graph** $I_{F\prec}$ for elimination order $\prec$ has an edge $X_i - X_j$ if X_i and X_j appear together in a factor generated by VE for elimination order $\prec$ on factors F



Elimination order:
{C,D,S,I,L,H,J,G}

every Xi eliminated connect uneliminated neighbors together!

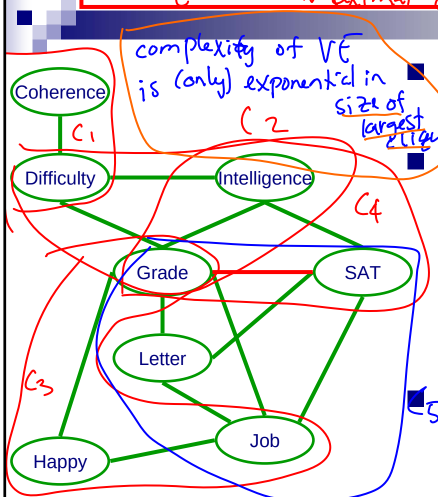


Induced graph and complexity of VE

Read complexity from cliques in induced graph

fully connected together
set of nodes
maximal clique: not subset of other cliques

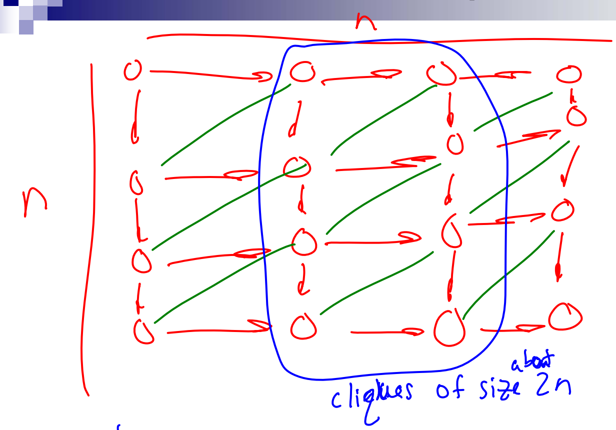
complexity of VE is (only) exponential in size of largest clique



Elimination order:
{C,D,I,S,L,H,J,G}

- Structure of induced graph encodes complexity of VE!!!
- **Theorem**: with ordering ≺
  - Every factor generated by VE subset of a maximal clique in $I_{F\prec}$
  - For every maximal clique in $I_{F\prec}$ corresponds to a factor generated by VE with ordering ≺
- **Induced width** (or treewidth)
  - Size of largest clique in $I_{F\prec}$ minus 1
  - *Minimal induced width* – induced width of best order $\prec$

Example: Large induced-width with small number of parents



total of n^2

only two parents

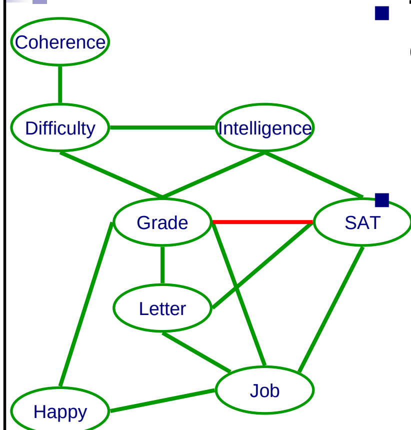
— moralization arcs

triangulate...
pick best order

cliques of size $O(n)$ $\Rightarrow$ inference with VE complexity $O(n^2 k^{2n})$

Compact representation $\not\Rightarrow$ Easy inference ☹

Finding optimal elimination order

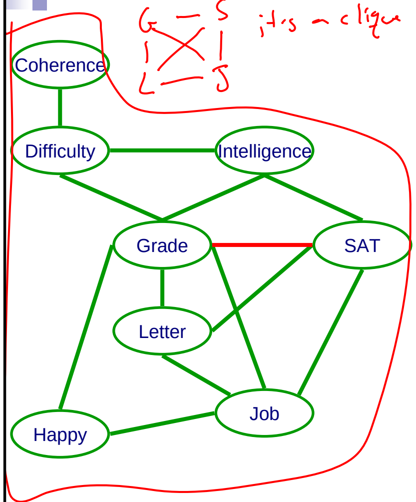


Elimination order: {C,D,I,S,L,H,J,G}

- **Theorem**: Finding best elimination order is NP-complete:
  - Decision problem: Given a graph, determine if there exists an elimination order that achieves induced width $\leq$ K
- **Interpretation**:
  - Hardness of finding elimination order in addition to hardness of inference
  - Actually, can find elimination order in time exponential in size of largest clique – same complexity as inference

Induced graphs and chordal graphs

it's a clique



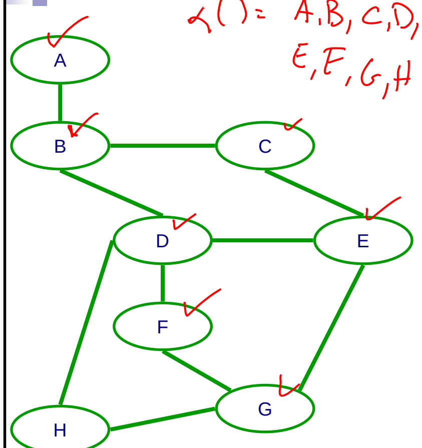
- **Chordal graph**:
  - Every cycle $X_1 - X_2 - \ldots - X_k - X_1$ with $k \geq 3$ has a chord
  - Edge $X_i - X_j$ for non-consecutive i & j
- **Theorem**:
  - Every induced graph is chordal
- "Optimal" elimination order easily obtained for chordal graph

Chordal graphs and triangulation

✓ Mark

last

$\prec(\ ) =$ A, B, C, D, E, F, G, H

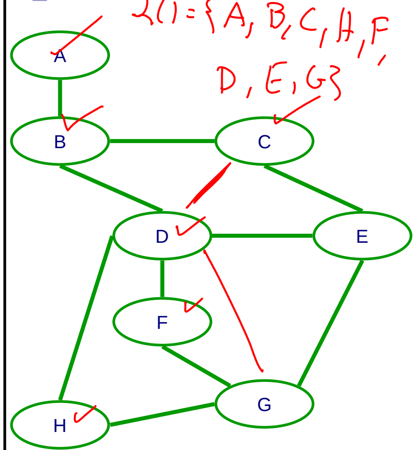


- **Triangulation**: turning graph into chordal graph
- **Max Cardinality Search**:
  - Simple heuristic
- Initialize unobserved nodes **X** as unmarked
- For k = |**X**| to 1
  - X ← unmarked var with most **marked** neighbors
  - $\prec$(X) ← k
  - Mark X
- **Theorem**: Obtains optimal order for chordal graphs
- Often, not so good in other graphs!

Minimum fill/size/weight heuristics

first

$\prec(\cdot) = \{A, B, C, H, F, D, E, G\}$



- Many more effective heuristics
  - see reading
- **Min (weighted) fill heuristic**
  - Often very effective
- Initialize unobserved nodes **X** as unmarked
- For k = 1 to |**X**|
  - X ← unmarked var whose elimination adds fewest edges
  - $\prec$(X) ← k
  - Mark X
  - Add fill edges introduced by eliminating X
- Weighted version:
  - Consider size of factor rather than number of edges

Choosing an elimination order

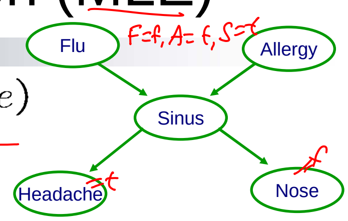
- Choosing best order is NP-complete
  - Reduction from MAX-Clique
- Many good heuristics (some with guarantees)
- Ultimately, can't beat NP-hardness of inference
  - Even optimal order can lead to exponential variable elimination computation
- In practice
  - Variable elimination often very effective
  - Many (many many) approximate inference approaches available when variable elimination too expensive
  - Most approximate inference approaches build on ideas from variable elimination

Announcements

- Recitation on advanced topic:
  - Carlos on Context-Specific Independence
  - On Monday Oct 16, 5:30-7:00pm in Wean Hall 4615A

o HW3 out later today

Most likely explanation (MLE)



- Query: $\underset{x_1,\dots,x_n}{\text{argmax}}\, P(x_1,\dots,x_n \mid e)$

- Using defn of conditional probs:

$$\underset{x_1,\dots,x_n}{\text{argmax}}\, P(x_1,\dots,x_n \mid e) = \underset{x_1,\dots,x_n}{\text{argmax}}\, \frac{P(x_1,\dots,x_n,e)}{P(e)}$$

constant

- Normalization irrelevant:

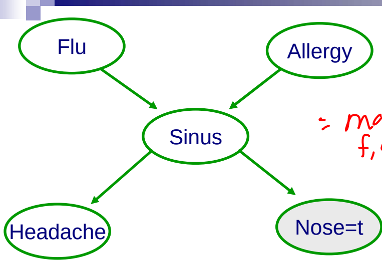
$$\underset{x_1,\dots,x_n}{\text{argmax}}\, P(x_1,\dots,x_n \mid e) = \underset{x_1,\dots,x_n}{\text{argmax}}\, P(x_1,\dots,x_n,e)$$

Max-marginalization

$\max_{f,s} P(f) \cdot P(s|f) \cdot P(N=t|s)$

$\max [P(F=t) \cdot P(s=t|F=t)\ P(N=t|S=t);$
$P(F=f) \cdot P(s=t \ldots;$
$P(F=t)\ P(s=f \ldots$
$P(F=f) \cdot P(s=f) \ldots]$

Example of variable elimination for MLE – Forward pass



$\max_{f,a,s,h} P(f) \cdot P(a)\ p(s|f,a) \cdot p(h|s)\ P(N=t|s)$

$= \max_{f,a,s} P(f)\ p(a)\ p(s|f,a) \cdot P(N=t|S)\ \underbrace{\max_h P(h|s)}_{g_1(s)}$

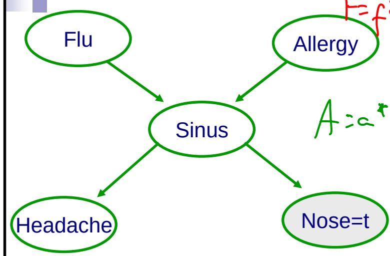
$= \max_{f,a} P(f) \cdot p(a) \cdot \underbrace{\max_s P(s|f,a) \cdot P(N=t|s)\, g_1(s)}_{g_2(f,a)}$

$= \max_f P(f) \cdot \underbrace{\max_a P(a) \cdot g_2(f,a)}_{g_3(f)}$

$g_4 = \max_f P(f) \cdot g_3(f)$

Interpretation:
$g_1(s)$ = highest prob h achieves for each value of S
$g_2(f,a)$ = highest prob. S & h | N=t for each value of f, a
g_4 = prob. MLE | N=t

Example of variable elimination for MLE – Backward pass



$F = f^* = \arg\max_f P(f) \cdot g_3(f)$

$A = a^* = \arg\max_a P(a) \cdot g_2(f^*, a)$

$S = s^* = \arg\max_s P(s \mid f^*, a^*) \cdot P(N=t \mid s) \, g_1(s)$

$H = h^* = \arg\max_h P(h \mid s^*)$

MLE Variable elimination algorithm – Forward pass

use max instead of sum
(no pruning)

- Given a BN and a MLE query $\max_{x_1,\ldots,x_n} P(x_1,\ldots,x_n,\mathbf{e})$
- Instantiate evidence $\mathbf{E}=\mathbf{e}$
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n, If $X_i \notin \mathbf{E}$
  - Collect factors $f_1,\ldots,f_k$ that include X_i
  - Generate a new factor by eliminating X_i from these factors

$$g = \max_{x_i} \prod_{j=1}^{k} f_j$$

  - Variable X_i has been eliminated!

cache g

MLE Variable elimination algorithm – Backward pass

- $\{x_1^*,\ldots, x_n^*\}$ will store maximizing assignment
- For i = n to 1, If $X_i \notin \mathbf{E}$ — reverse — f_j cannot depend on $X_1 \ldots X_{i-1}$ because they were eliminated before X_i
  - Take factors $f_1,\ldots,f_k$ used when X_i was eliminated
  - Instantiate $f_1,\ldots,f_k$, with $\{x_{i+1}^*,\ldots, x_n^*\}$
    - Now each f_j depends only on X_i
  - Generate maximizing assignment for X_i:

$$x_i^* \in \underbrace{\operatorname*{argmax}_{x_i} \prod_{j=1}^{k} f_j}$$

What you need to know about VE

- Variable elimination algorithm
  - Eliminate a variable:
    - Combine factors that include this var into single factor
    - Marginalize var from new factor
  - Cliques in induced graph correspond to factors generated by algorithm
  - Efficient algorithm ("only" exponential in induced-width, not number of variables)
    - If you hear: "Exact inference only efficient in tree graphical models"
    - You say: "No!!! Any graph with low induced width"
    - And then you say: "And even some with very large induced-width" (special recitation)
- Elimination order is important!
  - NP-complete problem
  - Many good heuristics
- Variable elimination for MLE
  - Only difference between probabilistic inference and MLE is "sum" versus "max"

What if I want to compute $P(X_i|x_0,x_{n+1})$ for each i?

$X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$

$x_0 = t$, $x_5 = f$

Compute: $P(X_i \mid x_0 = t, x_{n+1} = t)$

Variable elimination for each i? e.g., $X_1 \ldots X_{i-1}, X_{i+1} \ldots X_n$

eliminate X_1 $\quad g_1(X_2) = \sum_{x_1} P(x_0) \cdot P(x_1|x_0) \cdot P(X_2|x_1)$

Complexity of $P(X_i \mid x_0, x_{n+1}) = O(n)$

Variable elimination for every i, what's the complexity?

run VE n times:

naive $\underline{O(n^2)}$!

Reusing computation

$g_3(X_4) = \sum_{x_3} P(x_4|x_3) \cdot g_2(x_3)$

$g_2(X_3) = \sum_{x_2} P(X_3|x_2) \cdot g_1(x_2)$

$X_0 \rightarrow \cancel{X_1} \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$

$x_0 = t$, $x_5 = f$

"message" $g_1(X_2)$ m_{12}, $g_2(X_3)$ m_{23}, $g_3(X_4)$ m_{34}, $g_4(X_3)$; m_{21}, m_{32}, m_{43}

Compute: $P(X_i \mid x_0, x_{n+1})$

$P(X_3 \mid x_0, x_5)$ – compute $g_1(X_2), g_2(X_3)$, doesn't need $g_3(X_4)$

$P(X_4 \mid x_0, x_5)$ – compute $\underbrace{g_1(X_2), g_2(X_3), g_3(X_4)}_{\text{same } g_1}$

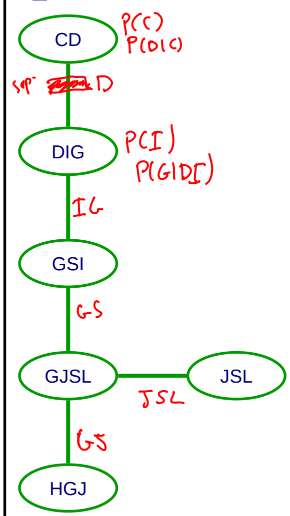
done!! :) $P(X_4|x_0, x_5) \propto g_3(X_4)\, P(X_5{=}f|X_4)$

→ need to eliminate X_4: done!! :)

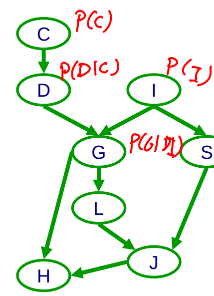
$g_4(X_3) = \sum_{x_4} P(X_5{=}f \mid X_4) \cdot P(X_4|X_3)$

compute each message once!!, so two passes $(O(n))$ gives you all probs.

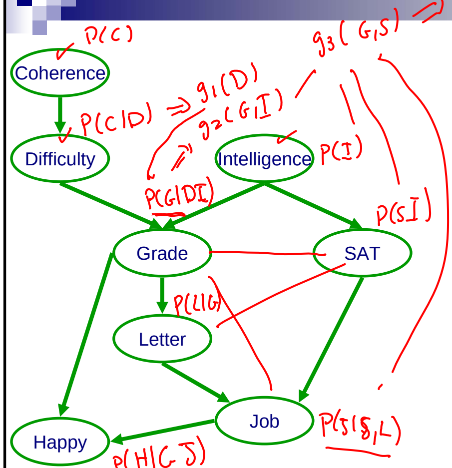
Cluster graph

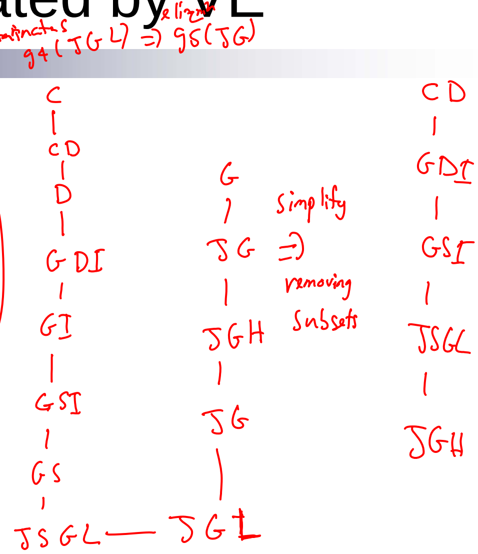


- **Cluster graph**: For set of factors *F*
  - Undirected graph
  - Each node i associated with a cluster $\mathbf{C}_i$
  - *Family preserving*: for each factor $f_j \in F$, $\exists$ node i such that $scope[f_j] \subseteq \mathbf{C}_i$
  - Each edge i – j is associated with a separator $\mathbf{S}_{ij} = \mathbf{C}_i \cap \mathbf{C}_j$



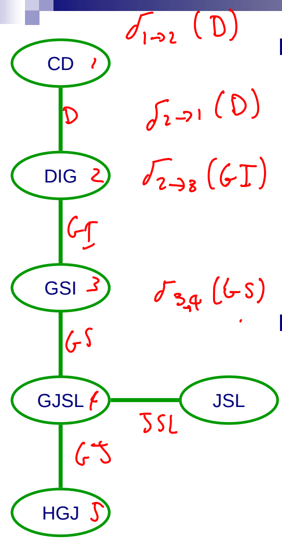
Factors generated by VE





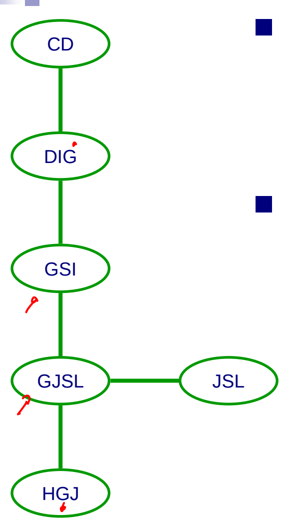
Elimination order:
{C,D,I,S,L,H,J,G}

Cluster graph for VE



- **VE generates cluster tree!**
  - One clique for each factor used/generated
  - Edge i – j, if f_i used to generate f_j
  - "Message" from i to j generated when marginalizing a variable from f_i
  - Tree because factors only used once
- **Proposition**:
  - "Message" δ_{ij} from i to j
  - Scope[δ_{ij}] $\subseteq$ $\mathbf{S}_{ij}$

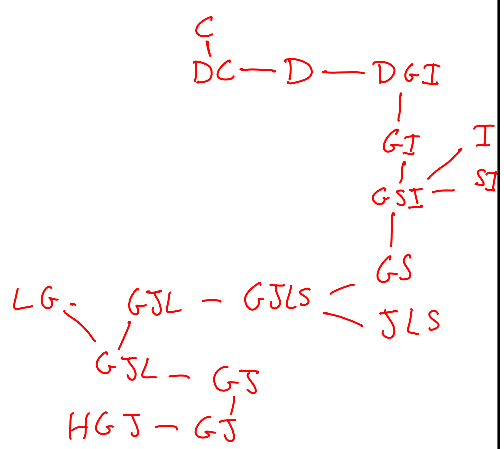
Running intersection property



- **Running intersection property (RIP)**
  - Cluster tree satisfies RIP if whenever X∈ $\mathbf{C}_i$ and X∈ $\mathbf{C}_j$ then X is in every cluster in the (unique) path from $\mathbf{C}_i$ to $\mathbf{C}_j$
- **Theorem**:
  - Cluster tree generated by VE satisfies RIP

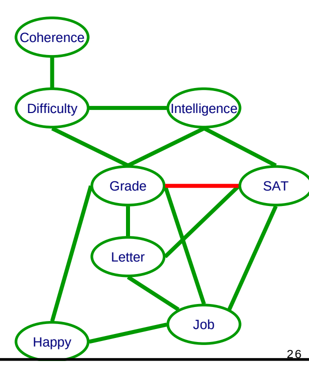
Constructing a clique tree from VE

- Select elimination order $\prec$
- Connect factors that would be generated if you run VE with order $\prec$
- Simplify!
  - Eliminate factor that is subset of neighbor

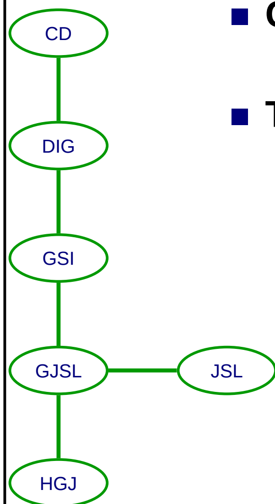


Find clique tree from chordal graph

- Triangulate moralized graph to obtain chordal graph
- Find maximal cliques
  - NP-complete in general
  - Easy for chordal graphs
  - Max-cardinality search
- Maximum spanning tree finds clique tree satisfying RIP!!!
  - Generate weighted graph over cliques
  - Edge weights (i,j) is separator size – $|C_i \cap C_j|$

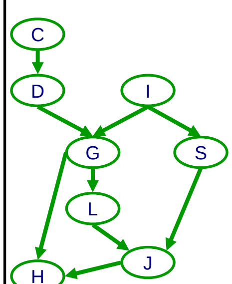


Clique tree & Independencies



- **Clique tree (or Junction tree)**
  - A cluster tree that satisfies the RIP
- **Theorem**:
  - Given some BN with structure *G* and factors *F*
  - For a clique tree *T* for *F* consider $\mathbf{C}_i - \mathbf{C}_j$ with separator $\mathbf{S}_{ij}$:
    - $\mathbf{X}$ – any set of vars in $\mathbf{C}_i$ side of the tree
    - $\mathbf{Y}$ – any set of vars in $\mathbf{C}_j$ side of the tree
  - Then, $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{S}_{ij})$ in BN
  - Furthermore, $I(T) \subseteq I(G)$

Variable elimination in a clique tree 1

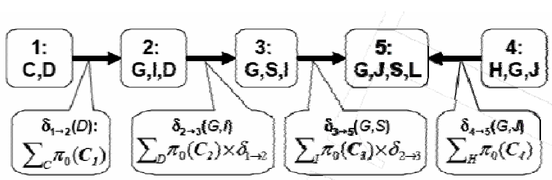


- **Clique tree for a BN**
  - Each CPT assigned to a clique
  - Initial potential $\pi_0(\mathbf{C}_i)$ is product of CPTs

Variable elimination in a clique tree 2

- **VE in clique tree to compute P(X_i)**
  - Pick a root (any node containing X_i)
  - Send messages recursively from leaves to root
    - Multiply incoming messages with initial potential
    - Marginalize vars that are not in separator
  - Clique *ready* if received messages from all neighbors

Belief from message

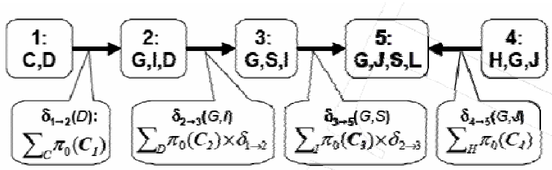


- **Theorem**: When **clique C_i is ready**
  - Received messages from all neighbors
  - Belief π_i(**C**$_i$) is product of initial factor with messages:

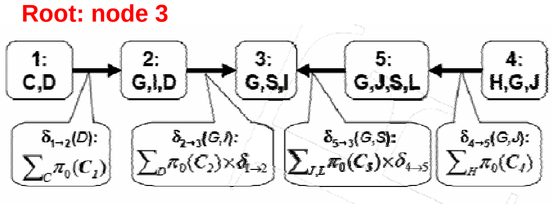
Choice of root

- **Message does not depend on root!!!**

Root: node 5

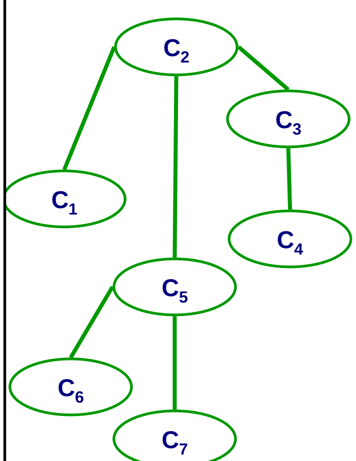


Root: node 3



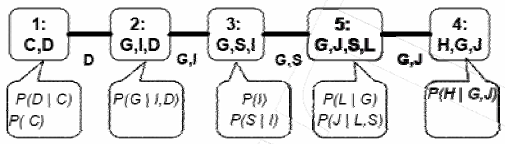
"Cache" computation: Obtain belief for all roots in linear time!!

Shafer-Shenoy Algorithm (a.k.a. VE in clique tree for all roots)



- Clique $\mathbf{C}_i$ *ready to transmit* to neighbor $\mathbf{C}_j$ if received messages from all neighbors but j
  - Leaves are always ready to transmit
- While $\exists$ $\mathbf{C}_i$ ready to transmit to $\mathbf{C}_j$
  - Send message $\delta_{i\to j}$
- Complexity: Linear in # cliques
  - One message sent each direction in each edge
- **Corollary**: At convergence
  - Every clique has correct belief

Calibrated Clique tree



- Initially, neighboring nodes don't agree on "distribution" over separators
- **Calibrated clique tree**:
  - At convergence, tree is *calibrated*
  - Neighboring nodes agree on distribution over separator

Answering queries with clique trees

- Query within clique
- Incremental updates – Observing evidence Z=z
  - Multiply some clique by indicator **1**(Z=z)
- Query outside clique
  - Use variable elimination!

Message passing with division

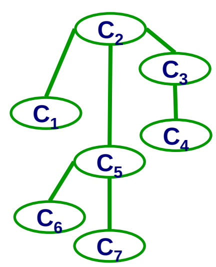
- Computing messages by multiplication:

- Computing messages by division:

Lauritzen-Spiegelhalter Algorithm (a.k.a. belief propagation)

Simplified description see reading for details

- Initialize all separator potentials to 1
  - $\mu_{ij} \leftarrow 1$
- All messages ready to transmit
- While $\exists\ \delta_{i\to j}$ ready to transmit
  - $\mu_{ij}' \leftarrow$
  - If $\mu_{ij}' \neq \mu_{ij}$
    - $\delta_{i\to j} \leftarrow$
    - $\pi_j \leftarrow \pi_j \times \delta_{i\to j}$
    - $\mu_{ij} \leftarrow \mu_{ij}'$
    - $\forall$ neighbors k of j, k$\neq$ i, $\delta_{j\to k}$ ready to transmit
- Complexity: Linear in # cliques
  - for the “right” schedule over edges (leaves to root, then root to leaves)
- **Corollary**: At convergence, every clique has correct belief



VE versus BP in clique trees

- VE messages (the one that multiplies)

- BP messages (the one that divides)

Clique tree invariant

- **Clique tree potential**:
  - Product of clique potentials divided by separators potentials

- Clique tree **invariant**:
  - P(**X**) = $\pi_{\mathcal{T}}(\mathbf{X})$

Belief propagation and clique tree invariant

- **Theorem**: Invariant is maintained by BP algorithm!
- BP reparameterizes clique potentials and separator potentials
  - At convergence, potentials and messages are marginal distributions

Subtree correctness

- **Informed message** from i to j, if all messages into i (other than from j) are informed
  - Recursive definition (leaves always send informed messages)
- **Informed subtree**:
  - All incoming messages informed
- **Theorem**:
  - Potential of connected informed subtree *T'* is marginal over scope[*T'*]
- **Corollary**:
  - At convergence, clique tree is *calibrated*
    - π_i = P(scope[π_i])
    - μ_{ij} = P(scope[μ_{ij}])

Clique trees versus VE

- Clique tree advantages
  - Multi-query settings
  - Incremental updates
  - Pre-computation makes complexity explicit

- Clique tree disadvantages
  - Space requirements – no factors are "deleted"
  - Slower for single query
  - Local structure in factors may be lost when they are multiplied together into initial clique potential

Clique tree summary

- Solve marginal queries for all variables in only twice the cost of query for one variable
- Cliques correspond to maximal cliques in induced graph
- Two message passing approaches
  - VE (the one that multiplies messages)
  - BP (the one that divides by old message)
- Clique tree invariant
  - Clique tree potential is always the same
  - We are only reparameterizing clique potentials
- Constructing clique tree for a BN
  - from elimination order
  - from triangulated (chordal) graph
- Running time (only) exponential in size of largest clique
  - Solve **exactly** problems with thousands (or millions, or more) of variables, and cliques with tens of nodes (or less)