Bayesian networks approximation

Eric Fabre

ANR StochMC, Feb. 13, 2014

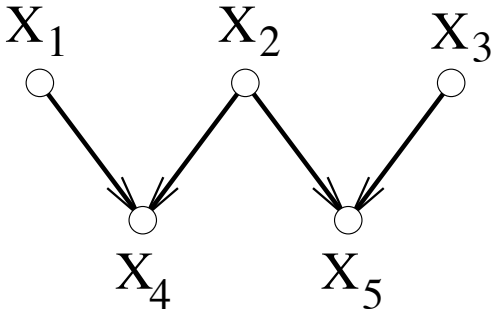
Outline

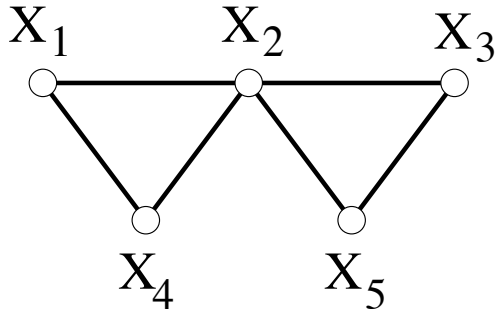
Motivation

Goal: simplify Bayes nets / Markov fields to make them tractable

Network of random variables

- $X_1, ..., X_n$ with $P_X = \Pi_i P_{X_i|\mathcal{P}(X_i)} = \Pi_i \phi(X_i, \mathcal{P}(X_i))$
- Ex. $P_X = P_{X_1} P_{X_2} P_{X_3} P_{X_4|X_1,X_2} P_{X_5|X_2,X_3}$





Inference: compute $P(X|Y = y)$ where Y is a subset of observed variables in X

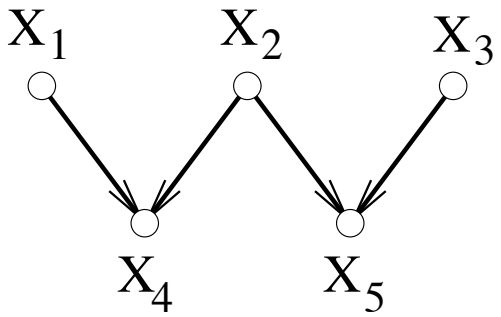
- tree structure $\Rightarrow$ inference is easy (linear)
- nb of cycles $\uparrow \Rightarrow$ complexity $\uparrow$

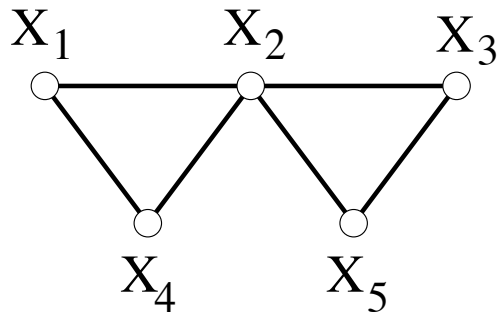
Motivation

Goal: simplify Bayes nets / Markov fields to make them tractable

Network of random variables

- $X_1, ..., X_n$ with $P_X = \Pi_i P_{X_i|\mathcal{P}(X_i)} = \Pi_i \phi(X_i, \mathcal{P}(X_i))$
- Ex. $P_X = P_{X_1} P_{X_2} P_{X_3} P_{X_4|X_1,X_2} P_{X_5|X_2,X_3}$



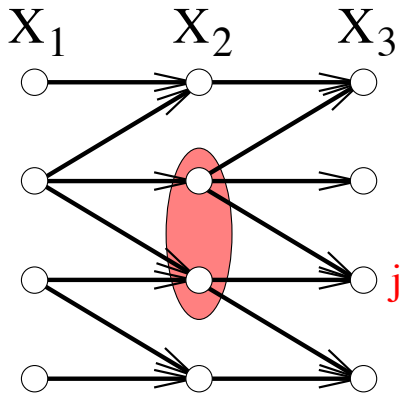


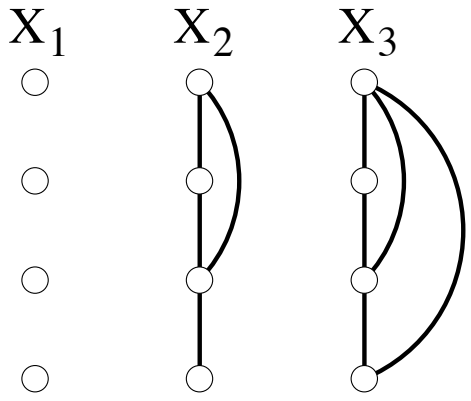
Inference: compute $P(X|Y = y)$ where Y is a subset of observed variables in X

- tree structure $\Rightarrow$ inference is easy (linear)
- nb of cycles $\uparrow \Rightarrow$ complexity $\uparrow$

Dynamic Bayesian networks

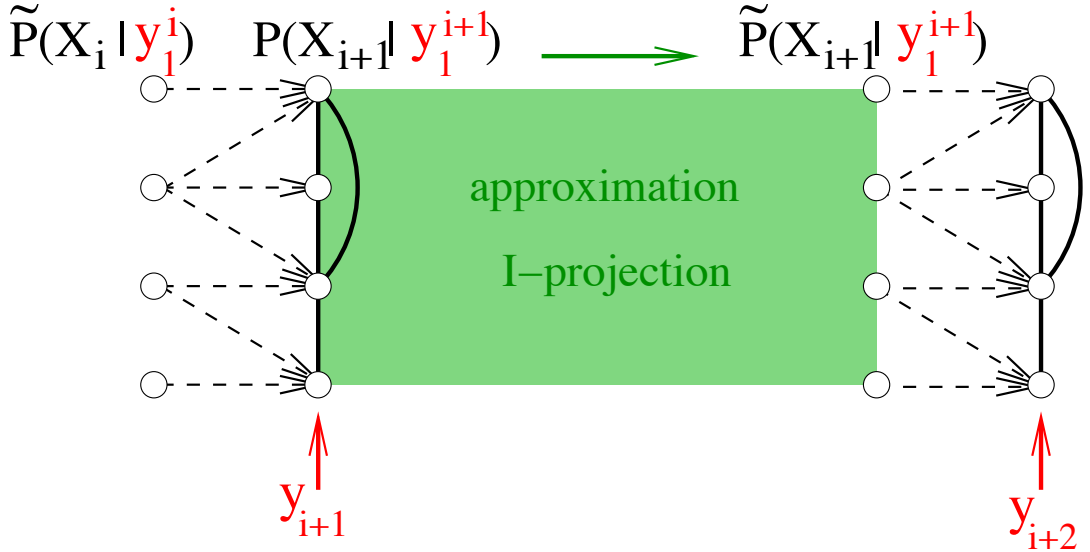
- $X_1, ..., X_n$ form a Markov chain, $P_X = P_{X_1} P_{X_2|X_1} P_{X_3|X_2} ...$
- each X_i itself is a large vector $X_i = [X_{i,j}]_{1 \leq j \leq m}$
- local dynamics:
  $P_{X_i|X_{i-1}} = \Pi_j P_{X_{i,j}|X_{i-1}}$ $\qquad$ $P_{X_{i,j}|X_{i-1}} = P_{X_{i,j}|X_{i-1,\mathcal{P}(j)}}$
- in marginals P_{X_i}, inner correlations increase as i grows
  this makes successive inferences $P_{X_i|y_1,...,y_i}$ tougher problems...





Factored frontier algorithm:

- approximate $P_{X_i|y_1,\ldots,y_i}$ by a simpler field (white noise)
  $\tilde{P}_{X_i|y_1,\ldots,y_i} = \Pi_j P_{X_{i,j}|y_1,\ldots,y_i}$
- then propagate to X_{i+1}, and incorporate new observation y_{i+1}

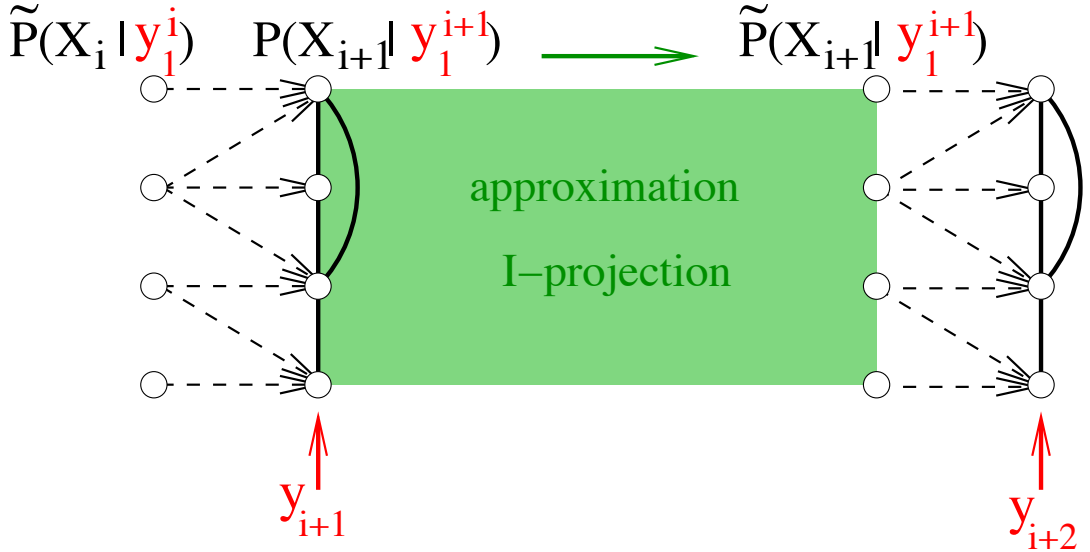


Two ways around complexity:

- run approximate inference on the exact *complex* model
- run exact inference on an approximate *simpler* model

Factored frontier algorithm:

- approximate $P_{X_i|y_1,\ldots,y_i}$ by a simpler field (white noise)
  $\tilde{P}_{X_i|y_1,\ldots,y_i} = \prod_j P_{X_{i,j}|y_1,\ldots,y_i}$
- then propagate to X_{i+1}, and incorporate new observation y_{i+1}



Two ways around complexity:

- run approximate inference on the exact *complex* model
- run exact inference on an approximate *simpler* model

Outline

Formalization

Idea: to simplify a network, remove edges one at a time

How ?

- An edge = a conditional independence test (yes/no)
- does not measure the *strength* of the link

Natural distance:

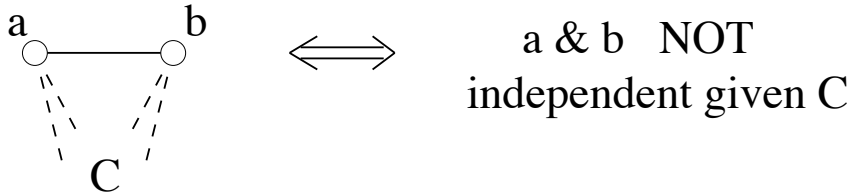
- Kullback-Leibler: $D(P_{A,B|C} \,\|\, P_{A|C} P_{B|C}) = I(A\,;B|C)$
- number of common "private" bits between A and B

Formalization

Idea: to simplify a network, remove edges one at a time

How ?

- An edge = a conditional independence test (yes/no)
- does not measure the *strength* of the link



Natural distance:

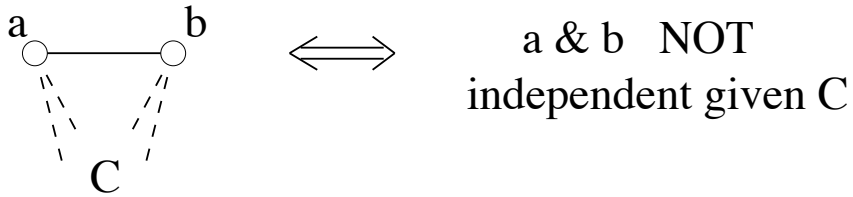
- Kullback-Leibler: $D(P_{A,B|C} \,\|\, P_{A|C} P_{B|C}) = I(A;B|C)$
- number of common "private" bits between A and B

Formalization

Idea: to simplify a network, remove edges one at a time

How ?

- An edge = a conditional independence test (yes/no)
- does not measure the *strength* of the link



Natural distance:

- Kullback-Leibler: $D(P_{A,B|C} \,\|\, P_{A|C} P_{B|C}) = I(A;B|C)$
- number of common "private" bits between A and B

Method

- given $P \sim \mathcal{G}$ with $\mathcal{G}$ a complex graph
  given $\mathcal{G}'$ a simpler graph
  find the best probability law Q such that $Q \sim \mathcal{G}'$

$$\min_Q D(P\|Q) = min_Q \sum_x p(x)\, log_2 \frac{p(x)}{q(x)}$$

- then optimize over graphs $\mathcal{G}'$

Wishes

- edge by edge simplification
- local cost of each edge
- additivity of costs

Method

- given $P \sim \mathcal{G}$ with $\mathcal{G}$ a complex graph
  given $\mathcal{G}'$ a simpler graph
  find the best probability law Q such that $Q \sim \mathcal{G}'$

$$\min_Q D(P\|Q) = min_Q \sum_x p(x)\, log_2 \frac{p(x)}{q(x)}$$

- then optimize over graphs $\mathcal{G}'$

Wishes

- edge by edge simplification
- local cost of each edge
- additivity of costs

General solution

Information geometry:

- assumptions:
  $\forall x,\ p(x) > 0,$ uniqueness of $Q,$ discrete values for X
- solution by I-projection (Csiszàr) over a log-linear space of distributions

Resolution

- IPFP (iterative proportional fitting procedure)
- Q obtained as a limit, and $D(P\|Q)$ is an infinite sum...
- Pythagora's theorem (additivity of distances)
- edges do not have a local cost
- edge by edge removal difficult

Triangulated graphs give all for free !

General solution

Information geometry:

- assumptions:
  $\forall x,\ p(x) > 0,$ uniqueness of $Q,$ discrete values for X
- solution by I-projection (Csiszàr) over a log-linear space of distributions

Resolution

- IPFP (iterative proportional fitting procedure)
- Q obtained as a limit, and $D(P\|Q)$ is an infinite sum...
- Pythagora's theorem (additivity of distances)
- edges do not have a local cost
- edge by edge removal difficult

Triangulated graphs give all for free !

General solution

Information geometry:

- assumptions:
  $\forall x,\ p(x) > 0,$ uniqueness of $Q,$ discrete values for X
- solution by I-projection (Csiszàr) over a log-linear space of distributions

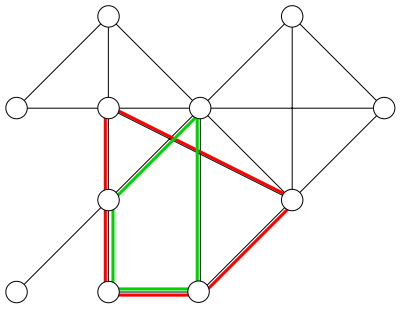
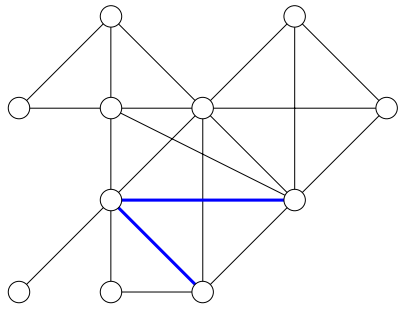
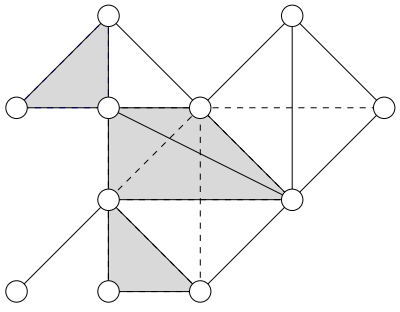
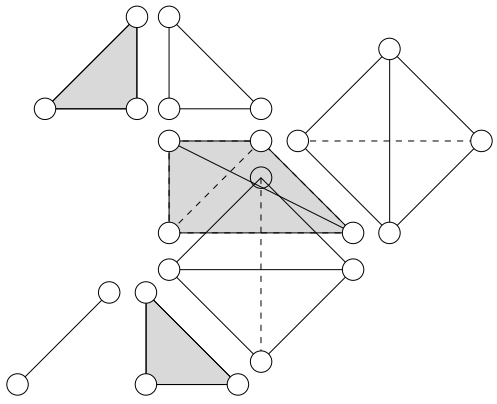
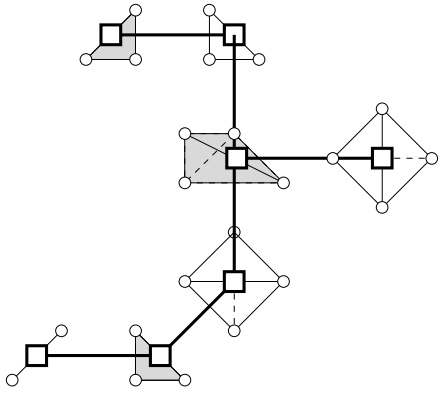
Resolution

- IPFP (iterative proportional fitting procedure)
- Q obtained as a limit, and $D(P\|Q)$ is an infinite sum...
- Pythagora's theorem (additivity of distances)
- edges do not have a local cost
- edge by edge removal difficult

Triangulated graphs give all for free !

Outline

Triangulated graphs generalize trees

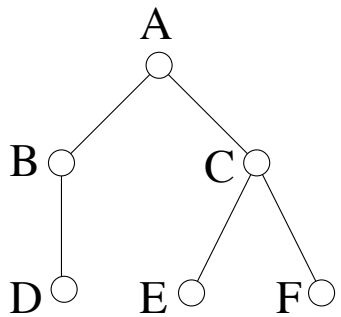






- tree-width of $\mathcal{G}$ = min over all triangulations $\mathcal{T}$ of $\mathcal{G}$ of the largest clique in $\mathcal{T}$
- related to the junction tree construction

Coding theorem

Theorem

$Q \sim \mathcal{T}$ *and* $\mathcal{T} = (V, E) =$ *tree*
then $Q \Leftrightarrow \{Q_{A,B} : (A, B) \in E\}$

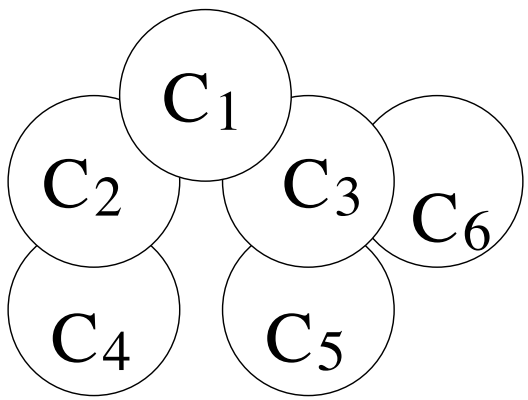


$$Q_{A,\ldots,F} = Q_A \; Q_{B|A} \; Q_{C|A} \; Q_{D|B} \; Q_{E|C} \; Q_{F|C}$$

Coding theorem (2)

Theorem

$Q \sim \mathcal{G}$ and $\mathcal{G}$ triangulated graph
then $Q \Leftrightarrow \{Q_C : C$ maximal clique in $\mathcal{G}\}$



$$Q = Q_{C_1}\ Q_{C_2 \ominus C_1 | C_2 \cap C_1}\ Q_{C_3 \ominus C_1 | C_3 \cap C_1} \cdots$$

I-projection

- one always has
  $D(P_{X,Y} \,\|\, Q_{X,Y}) = D(P_X \,\|\, Q_X) + D(P_{Y|X} \,\|\, Q_{Y|X})$
- $P \sim \mathcal{G}$ with target graph $\mathcal{G}'$ triangulated
  let $Q \sim \mathcal{G}'$ and let C be a maximal clique in $\mathcal{G}'$

$$D(P\|Q) = D(P_C \,\|\, Q_C) + D(P_{rest|C} \,\|\, Q_{rest|C})$$

  if $Q \sim \mathcal{G}'$ minimizes the distance, then $Q_C \equiv P_C$
- Q is then defined by $\{Q_C \triangleq P_C \,:\, C \text{ maximal clique in } \mathcal{G}'\}$

Properties:

- unique solution
- direct computation of Q
- no assumption on P

I-projection

- one always has
  $D(P_{X,Y} \,\|\, Q_{X,Y}) = D(P_X \,\|\, Q_X) + D(P_{Y|X} \,\|\, Q_{Y|X})$
- $P \sim \mathcal{G}$ with target graph $\mathcal{G}'$ triangulated
  let $Q \sim \mathcal{G}'$ and let C be a maximal clique in $\mathcal{G}'$

$$D(P\|Q) = D(P_C \,\|\, Q_C) + D(P_{rest|C} \,\|\, Q_{rest|C})$$

  if $Q \sim \mathcal{G}'$ minimizes the distance, then $Q_C \equiv P_C$
- Q is then defined by $\{Q_C \triangleq P_C \ : \ C \text{ maximal clique in } \mathcal{G}'\}$

Properties:

- unique solution
- direct computation of Q
- no assumption on P

I-projection

- one always has
  $D(P_{X,Y} \,\|\, Q_{X,Y}) = D(P_X \,\|\, Q_X) + D(P_{Y|X} \,\|\, Q_{Y|X})$
- $P \sim \mathcal{G}$ with target graph $\mathcal{G}'$ triangulated
  let $Q \sim \mathcal{G}'$ and let C be a maximal clique in $\mathcal{G}'$

$$D(P\|Q) = D(P_C \,\|\, Q_C) + D(P_{rest|C} \,\|\, Q_{rest|C})$$

  if $Q \sim \mathcal{G}'$ minimizes the distance, then $Q_C \equiv P_C$
- Q is then defined by $\{Q_C \triangleq P_C \ : \ C \text{ maximal clique in } \mathcal{G}'\}$

Properties:

- unique solution
- direct computation of Q
- no assumption on P

I-projection

- one always has
  $D(P_{X,Y} \,\|\, Q_{X,Y}) = D(P_X \,\|\, Q_X) + D(P_{Y|X} \,\|\, Q_{Y|X})$
- $P \sim \mathcal{G}$ with target graph $\mathcal{G}'$ triangulated
  let $Q \sim \mathcal{G}'$ and let C be a maximal clique in $\mathcal{G}'$

$$D(P\|Q) = D(P_C \,\|\, Q_C) + D(P_{rest|C} \,\|\, Q_{rest|C})$$

  if $Q \sim \mathcal{G}'$ minimizes the distance, then $Q_C \equiv P_C$
- Q is then defined by $\{Q_C \triangleq P_C \ : \ C \text{ maximal clique in } \mathcal{G}'\}$

Properties:

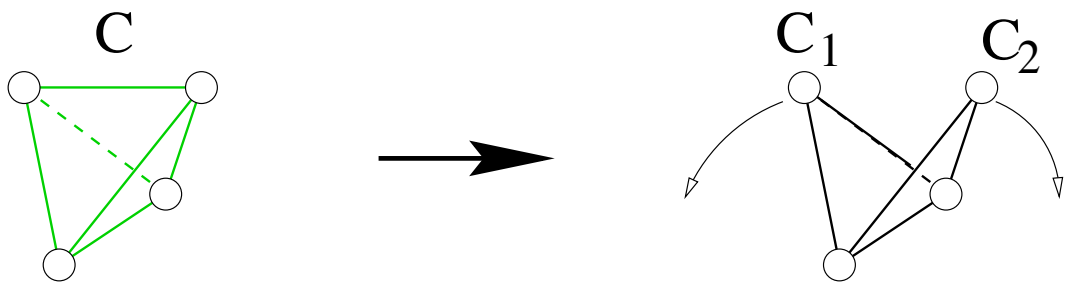
- unique solution
- direct computation of Q
- no assumption on P

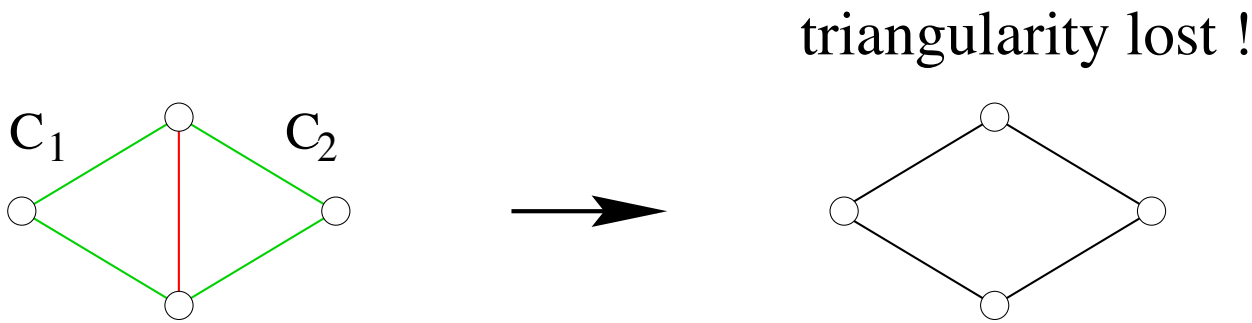
Surgery

Question: how to remove a single edge to a triangulated graph ?

Theorem

$\mathcal{G}$ triangulated graph, $\mathcal{G}' = \mathcal{G} \ominus (A, B)$ is triangulated iff edge (A, B) in $\mathcal{G}$ is a **green edge**,
i.e. belongs to a unique maximal clique of $\mathcal{G}$.

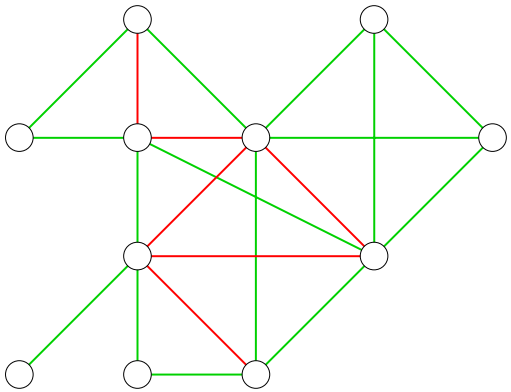


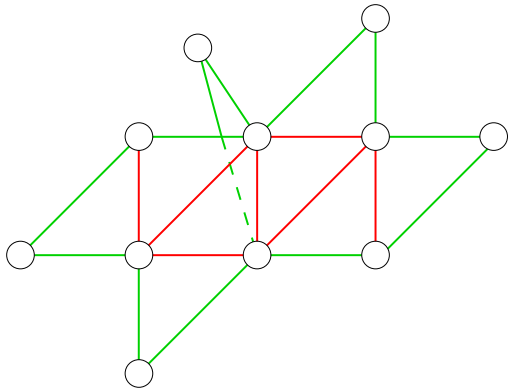


Green edges

Q: Are there many green edges ?

R: yes! they form the "skin" of the triangulated graph.





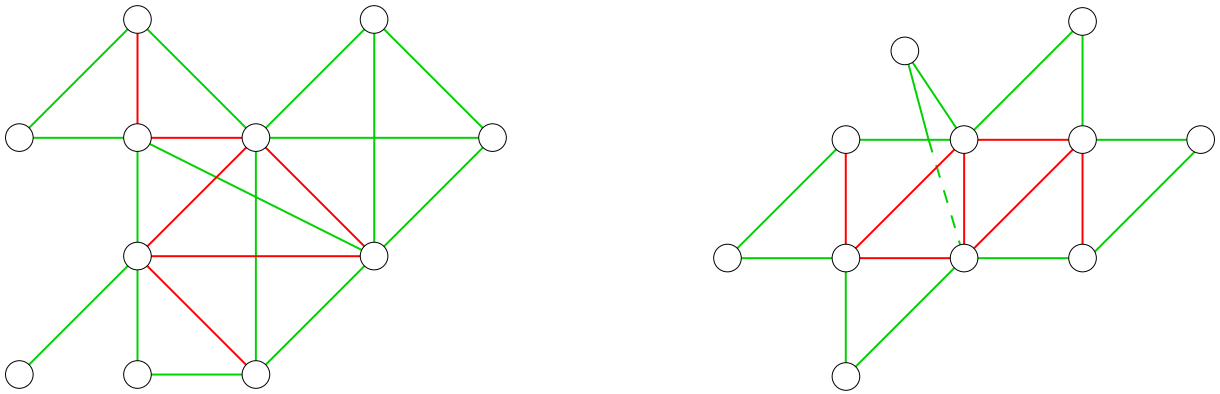
Properties

- at least 2 green edges attached to each node of degree ≥ 2
- a green edge is either separating (isthmus) or belongs to a green cycle
- $\exists$ green path between any two nodes
- $\exists$ green cycle containing any two nodes that are not separated by an isthmus

Green edges

Q: Are there many green edges ?

R: yes! they form the "skin" of the triangulated graph.



Properties

- at least 2 green edges attached to each node of degree ≥ 2
- a green edge is either separating (isthmus) or belongs to a green cycle
- $\exists$ green path between any two nodes
- $\exists$ green cycle containing any two nodes that are not separated by an isthmus

Green edges (2)

Theorem

Let $\mathcal{G} \supset \mathcal{G}'$ be triangulated graphs,
there exists a decreasing sequence of triangulated graphs

$$\mathcal{G} = \mathcal{G}_0 \supset \mathcal{G}_1 \supset \mathcal{G}_2 \supset \ldots \supset \mathcal{G}_n = \mathcal{G}'$$

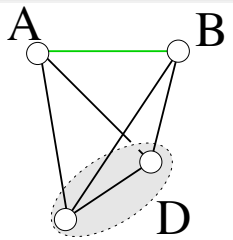
such that $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ differ by a single (green) edge.

Outline

Additivity of distances

Theorem

Let $\mathcal{G} \supset \mathcal{G}' \supset \mathcal{G}''$ be triangulated graphs,
and $P \sim \mathcal{G}$, $Q \sim \mathcal{G}'$, $R \sim \mathcal{G}''$ resp. best approximations of P,
then $D(P\|R) = D(P\|Q) + D(Q\|R)$



clique C

Proof. assume wlog $\mathcal{G}' = \mathcal{G} \ominus (A, B)$

$$\begin{aligned} P &= P_{A,B|D} \quad P_D \; P_{rest|C} \quad \mathcal{G} \\ Q &= P_{A|D} \; P_{B|D} \quad P_D \; P_{rest|C} \quad \mathcal{G}' \\ R &= R_{A|D} \; R_{B|D} \quad R_D \; R_{rest|C} \quad \mathcal{G}'' \end{aligned}$$

$$\begin{aligned} D(P_{A,B|D} \,\|\, R_{A|D} R_{B|D}) &= D(P_{A,B|D} \,\|\, P_{A|D} P_{B|D}) \\ &\quad + D(P_{A|D} P_{B|D} \,\|\, R_{A|D} R_{B|D}) \end{aligned}$$

Corollary

$\mathcal{G}' = \mathcal{G} \ominus (A, B)$, $P \sim \mathcal{G}$, $Q \sim \mathcal{G}'$ best approximation of P on $\mathcal{G}'$,

$$D(P\|Q) = D(P_{A,B|C} \,\|\, P_{A|C}\, P_{B|C}) = I(A; B|C)$$

where C is the (unique) maximal clique containing edge (A, B) in $\mathcal{G}$.

- involves P only on the (unique) clique C containing edge (A, B) : locality of the cost
- in a decreasing sequence $\mathcal{G} = \mathcal{G}_0 \supset \mathcal{G}_1 \supset \ldots$ of triangulated graphs, the distance computation always involves the initial probability P (on $\mathcal{G}$)

Corollary

$\mathcal{G}' = \mathcal{G} \ominus (A, B)$, $P \sim \mathcal{G}$, $Q \sim \mathcal{G}'$ *best approximation of* P *on* $\mathcal{G}'$,

$$D(P\|Q) = D(P_{A,B|C} \,\|\, P_{A|C}\, P_{B|C}) = I(A; B|C)$$

where C *is the (unique) maximal clique containing edge* (A, B) *in* $\mathcal{G}$.

- involves P only on the (unique) clique C containing edge (A, B) : locality of the cost
- in a decreasing sequence $\mathcal{G} = \mathcal{G}_0 \supset \mathcal{G}_1 \supset \ldots$ of triangulated graphs, the distance computation always involves the initial probability P (on $\mathcal{G}$)

Distance to white noise

White noise:

- $\mathcal{W} =$ graph with no edge (still same nodes as $\mathcal{G}$)
- $P \sim \mathcal{G}$, best probability $I \sim \mathcal{W}$ satisfies $I = \prod_{S \in V} P_S$
- $D(P\|I)$ can be computed by additivity through any decreasing sequence $\mathcal{G} = \mathcal{G}_0 \supset \mathcal{G}_1 \supset \ldots \supset \mathcal{G}_n = \mathcal{W}$

Distance to white noise

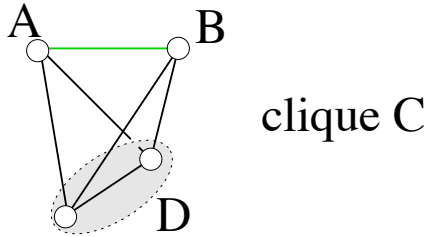
White noise:

- $\mathcal{W} =$ graph with no edge (still same nodes as $\mathcal{G}$)
- $P \sim \mathcal{G}$, best probability $I \sim \mathcal{W}$ satisfies $I = \prod_{S \in V} P_S$
- $D(P\|I)$ can be computed by additivity through any decreasing sequence $\mathcal{G} = \mathcal{G}_0 \supset \mathcal{G}_1 \supset ... \supset \mathcal{G}_n = \mathcal{W}$

Theorem

There exists weights $w_P(D)$ associated to cliques D of $\mathcal{G}$ such that

$$D(P\|I) = \sum_{D \text{ clique in } \mathcal{G}} w_P(D)$$



Proof. Remove (A, B) in clique C: $D(P\|I) = D(P\|Q) + D(Q\|I)$
If the theorem holds, one has

$$D(P\|Q) = I(A; B|D) = \sum_{E \subseteq D} w_P(E \cup \{A, B\})$$

By the Moëbius transform, one gets:

$$w_P(E \cup \{A, B\}) = \sum_{E \subseteq D} (-1)^{|D-E|} I(A; B|E)$$

Theorem

There exists weights $w_P(D)$ associated to cliques D of $\mathcal{G}$ such that

$$D(P\|I) = \sum_{D \text{ clique in } \mathcal{G}} w_P(D)$$

sum over all (non necessarily maximal) cliques D of $\mathcal{G}$

Examples

- $w_P(\emptyset) = 0$
- $w_P(\{A\}) = 0$
- $w_P(\{A, B\}) = I(A; B)$
- $w_P(\{A, B, C\}) = I(A; B|C) - I(A; B)$ sym in A, B, C
- $w_P(D)$ can be ≥ 0 or ≤ 0 for $|D| \geq 3$

Outline

Best triangulated graph

Remark:

- if Q best approximation of P on triangulated graph $\mathcal{G}'$, then

$$D(P\|I) = D(P\|Q) + D(Q\|I)$$

- so $\min_{\mathcal{G}'} D(P\|Q) \quad \Leftrightarrow \quad \max_{\mathcal{G}'} D(Q\|I)$

Hierarchy of triangulated graphs:

- $\mathcal{T}_p$ = triangulated graphs over vertices V, where cliques have at most p nodes

- $p \uparrow \Rightarrow \uparrow$nb of edges $\Rightarrow$ Q closer to P (further away from I)

Best triangulated graph

Remark:

- if Q best approximation of P on triangulated graph $\mathcal{G}'$, then

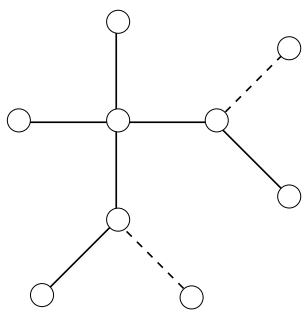
$$D(P\|I) = D(P\|Q) + D(Q\|I)$$

- so $\min_{\mathcal{G}'} D(P\|Q) \quad \Leftrightarrow \quad \max_{\mathcal{G}'} D(Q\|I)$

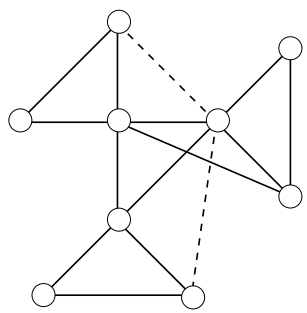
Hierarchy of triangulated graphs:

- $\mathcal{T}_p$ = triangulated graphs over vertices V, where cliques have at most p nodes

TO 2



TO 3



- $p \uparrow \Rightarrow \uparrow$nb of edges $\Rightarrow$ Q closer to P (further away from I)

Greedy algorithms

Best tree approximation:

$$\max_{\mathcal{G}' \in \mathcal{T}_2} D(Q\|I) \;=\; \max_{\mathcal{G}' \in \mathcal{T}_2} \sum_{\text{edge } \{A,B\} \in \mathcal{G}'} I(A;B)$$

- a best covering tree problem: greedy algo
- already discovered by [Chow et al., '68] !

Best $\mathcal{T}_p$ approximation:

$$\max_{\mathcal{G}' \in \mathcal{T}_p} \sum_{\text{edge } \{A,B\} \in \mathcal{G}'} I(A;B) + \sum_{\text{clique } \{A,B,C\} \in \mathcal{G}'} w_P(\{A,B,C\}) + \ldots$$

- greedy algos are sub-optimal, but not so bad [Malvestuto, '91]

Greedy algorithms

Best tree approximation:

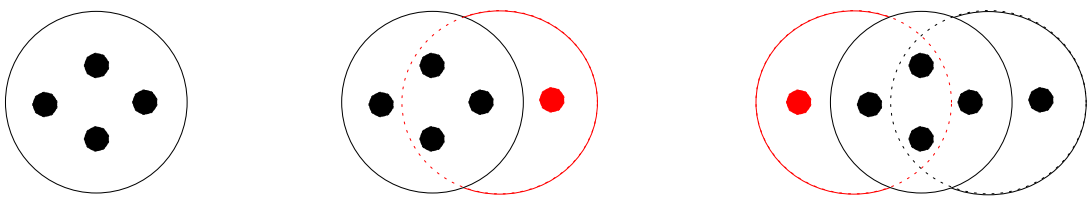
$$\max_{\mathcal{G}' \in \mathcal{T}_2} D(Q\|I) \;=\; \max_{\mathcal{G}' \in \mathcal{T}_2} \sum_{\text{edge } \{A,B\} \in \mathcal{G}'} I(A;B)$$

- a best covering tree problem: greedy algo
- already discovered by [Chow et al., '68] !

Best $\mathcal{T}_p$ approximation:

$$\max_{\mathcal{G}' \in \mathcal{T}_p} \sum_{\text{edge } \{A,B\} \in \mathcal{G}'} I(A;B) + \sum_{\text{clique } \{A,B,C\} \in \mathcal{G}'} w_P(\{A,B,C\}) + \ldots$$

- greedy algos are sub-optimal, but not so bad [Malvestuto, '91]



Conclusion

Summary

- Idea : simplify the model, then apply an exact algorithm
- Bayesian networks : easy with triangulated graphs

Questions

- Link between $D(P\|Q)$ and the quality of estimators built from Q instead of P ?
- Of interest to Blaise's problems ?
- What about networks of *dynamic* (probabilistic) systems ?

Conclusion

Summary

- Idea : simplify the model, then apply an exact algorithm
- Bayesian networks : easy with triangulated graphs

Questions

- Link between $D(P\|Q)$ and the quality of estimators built from Q instead of P ?
- Of interest to Blaise's problems ?
- What about networks of *dynamic* (probabilistic) systems ?