

Introduction to IAMB, GLL & LGL

BMIF 330, Spring 2008

Constantin Aliferis

Main points

- Learning causality is important & hard.
- Learning causality with Bayesian networks is well developed but until recently non-scalable.
- Localized learning is a very important approach to scale up causal learning & achieve optimal feature selection
- IAMB: a family of algorithms for learning the Markov Blanket of a variable
- GLL: A generalized framework for local learning.
- HITON and MMPC as instantiations of GLL.
- LGL: A general framework for Locally- constrained Global learning.
- MMHC as an instantiation of LGL.
- Indicative experimental results showing promise of framework.

Problem: Computational and sample complexity

- Learning Bayesian Networks and Causal Probabilistic Models is known to be worst-case intractable.
- Often very large samples are also required.
- We need techniques to scale-up learning to handle thousands or even millions of variables (with as small sample as required by function to be learnt).

Problem: Computational and sample complexity

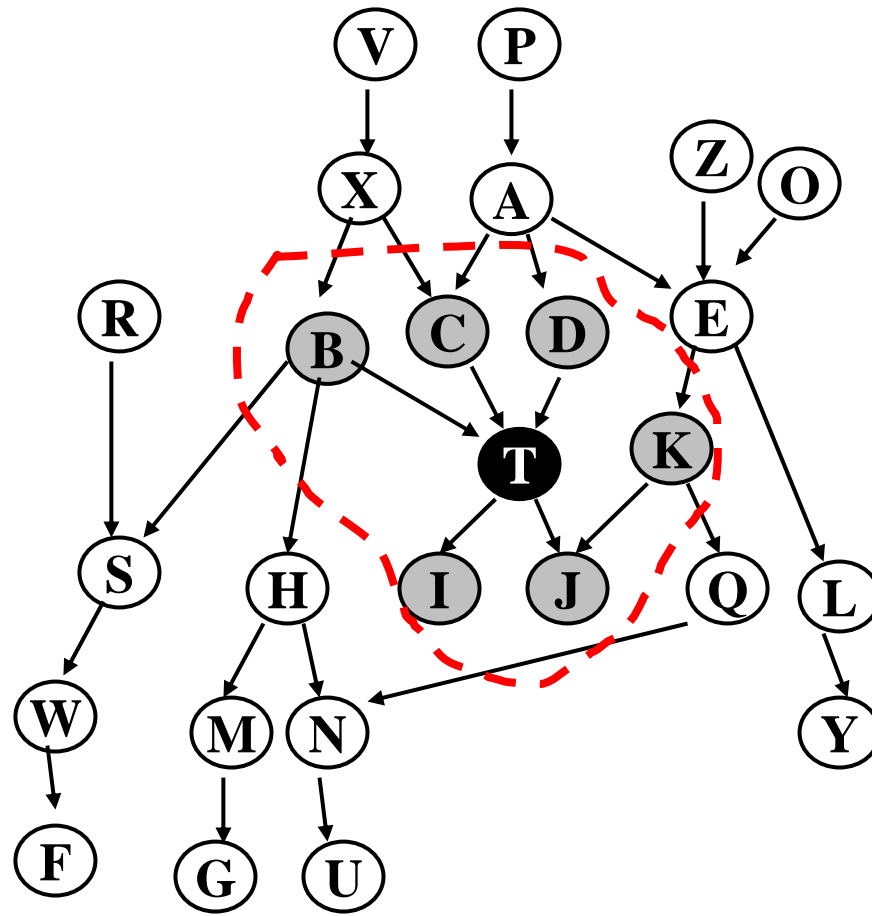
- When we started working on bioinformatics applications of BN-based causal discovery (in 2000) predictions were bleak:

“In our view, inferring complete causal models (i.e., causal Bayesian Networks) is essentially impossible in large-scale data mining applications with thousands of variables”. Silverstein, Brin, Motwani, Ullman. Data Mining and Knowledge Discovery, July 2000, pp. 163-192.

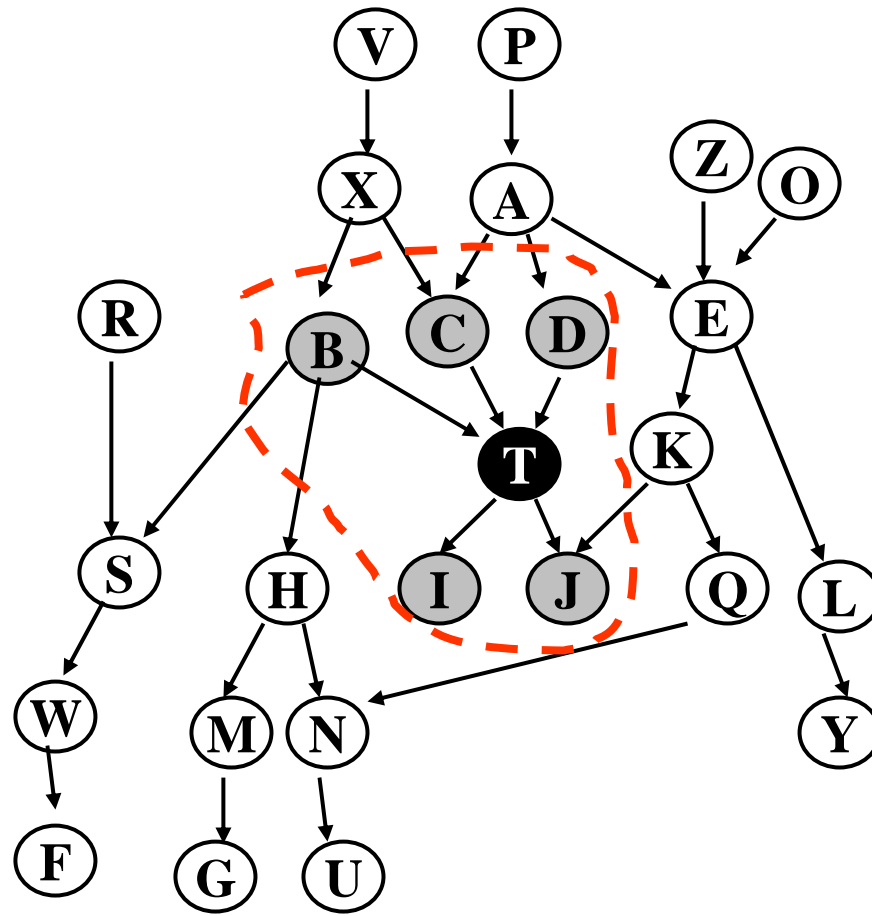
Strategies for Scaling up Causal Discovery

1. Develop algorithms with **good average-case performance**. Such algorithms would be tractable for many real-life datasets.
2. Abandon the effort to learn the full causal graph and instead develop methods that **learn the local neighborhood of a specific variable** directly.
3. Abandon the effort to learn the fully oriented causal graph and instead develop methods that **learn the unoriented graph**.
4. **Constrain the possible relationships** among variables and then learn the full causal graph.
5. Learn the full graph but **focus on very special types of distributions** (e.g., Naïve-Bayes distributions or tree-like graphs).
6. Abandon the effort to learn the full causal graph and instead develop methods that **find a portion of the true arcs** (not specific to some target variable).

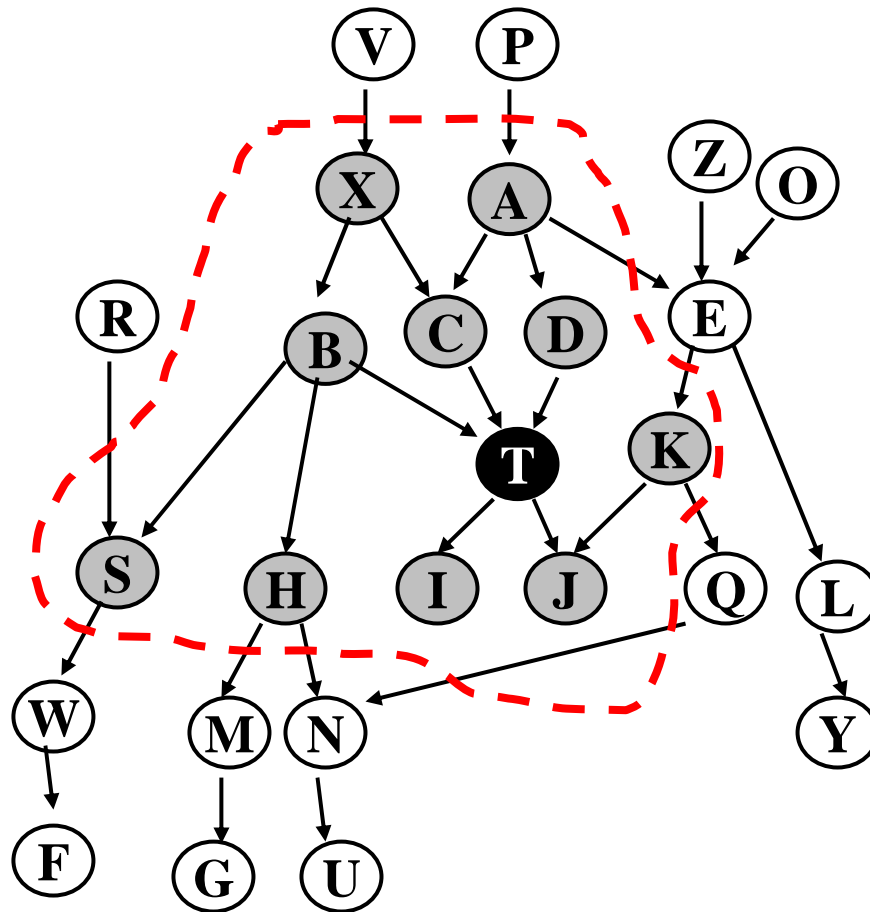
Problem #1: Consider a target variable T and discover Markov Blanket of T



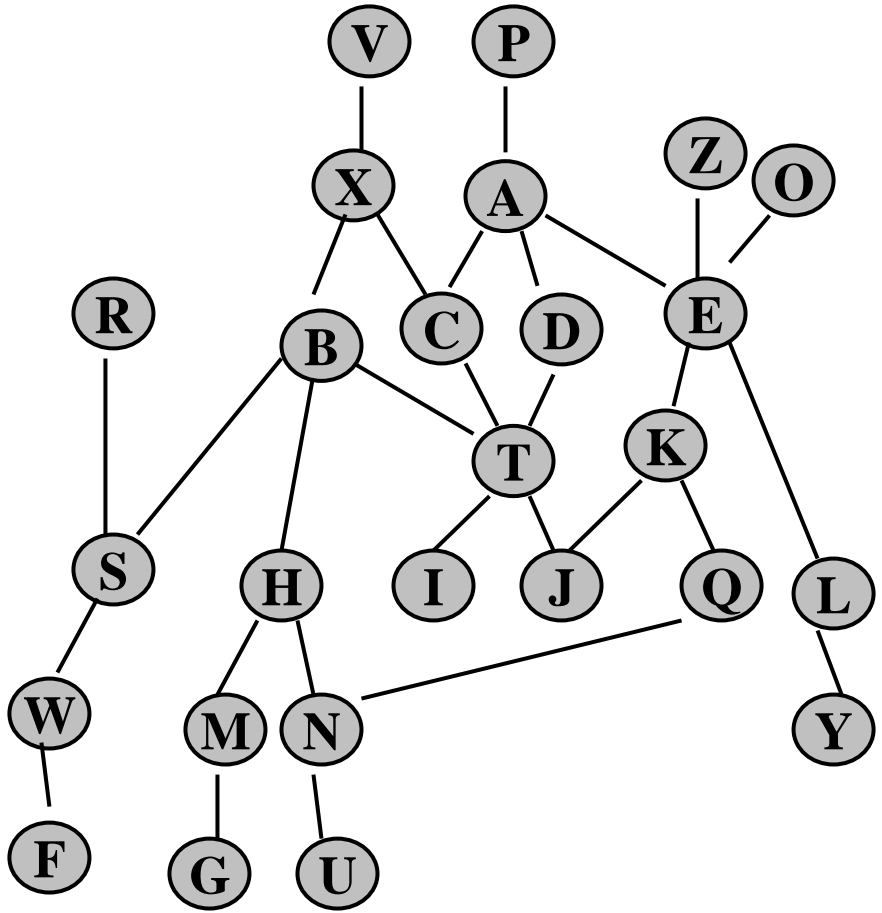
Problem #2: Consider a target variable T and Parents and Children of T



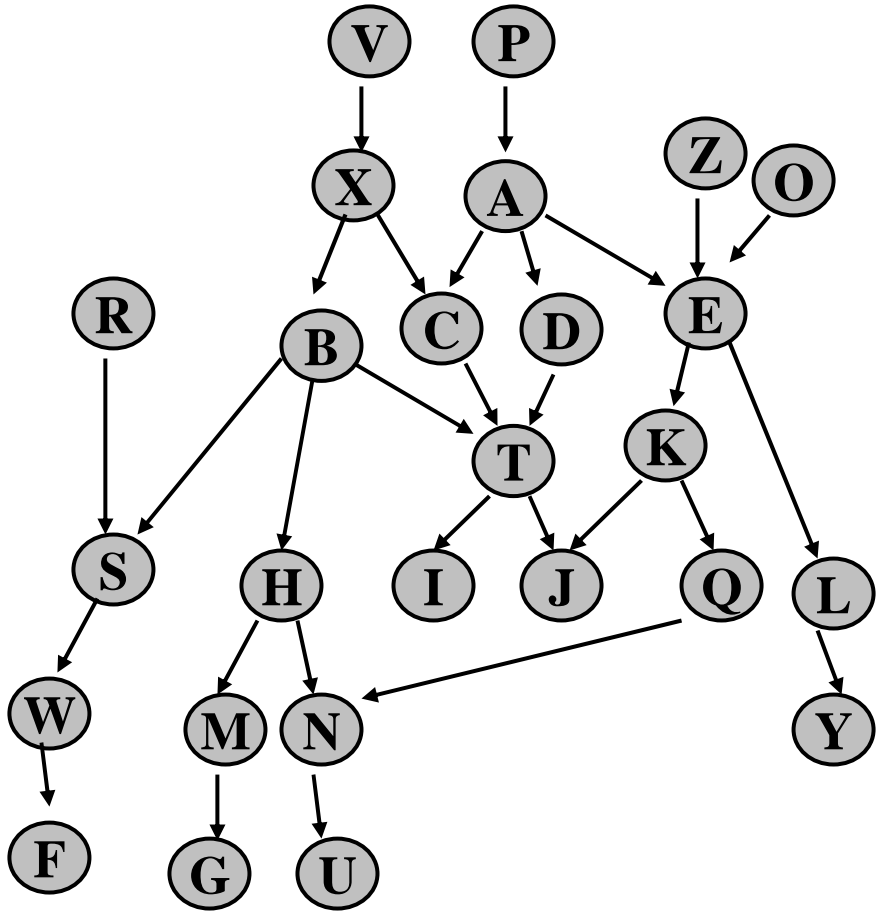
Problem #3: Consider a target variable T and Parents and discover Regions (say, of depth 2 edges) around T



Problem #4: Discover undirected graph



Problem #5: Discover directed graph



Many more discovery
questions/algorithm output not
mentioned here

Important observations about Markov Blanket(s)

- **Definition:** The MB of variable T (denoted by $MB(T)$) is a set that renders all variables (except T and $MB(T)$) independent of T conditioned on $MB(T)$
- **“Padding” a MB:** If $M1$ is a $MB(T)$ then $M2=\{M1,X\}$ (where X is a variable not in $M1$) is also a $MB(T)$.
- **Non-redundancy:** We are interested in $MB(T)$ that has not redundant variables, i.e., if I remove one or more variables from it, it stops being a MB.
- **Minimality:** In some distributions there are many non-reducible MBs. Then the minimal and non-reducible $MB(T)$ is often of interest for FS.
- **Uniqueness:** In faithful distributions, there is a unique non reducible (and thus minimal) $MB(T)$. We will focus on discovery of this $MB(T)$ in faithful distributions.
- **Causality:** In faithful distributions with causal sufficiency, $MB(T)$ contains the direct causes, direct effects and direct causes of direct effects of T . This provides a **graphical definition** for the minimal (or equivalently in this case, non-reducible) $MB(T)$.
- **Relevancy:** In faithful distributions, $MB(T)$ members are Kohavi-John strongly relevant, variables without a path connecting to some $MB()$ member(s) are KJ-irrelevant, and variables with a path connecting to some $MB()$ member(s) are KJ-weakly relevant.
- **Optimal predictor set:** In faithful distributions, $MB(T)$ is the smallest set that gives maximum predictivity for T assuming classifiers that can learn any function and loss functions that require effectively learning the conditional distribution: **$P(T | \text{predictor variable set})$** . This is the **solution to the “standard” feature selection problem**.

IAMB: when sample is large relative to $|MB(T)|$

Iterative Associative Markov Blanket (IAMB)

Input: - dataset D , - target variable T ,

Output: $MB(T)$

Start with an empty $CurrentMB$

Phase I:

Repeat

Find the variable V_i that maximizes $f(V_i ; T | CurrentMB)$

// f returns a non-zero value for every variable that is a member of the Markov Blanket; Typically a measure of association appropriate for the distribution of D // ← subtle point here

If not $I(V_i ; T | CurrentMB)$ Admit candidate variable V_i into $CurrentMB$,

Else Exit Loop

Until False

Phase II:

For all members V_j of $CurrentMB$

Eliminate V_j from MB if $I(V_j ; T | CurrentMB - \{V_j\})$

Return $CurrentMB$

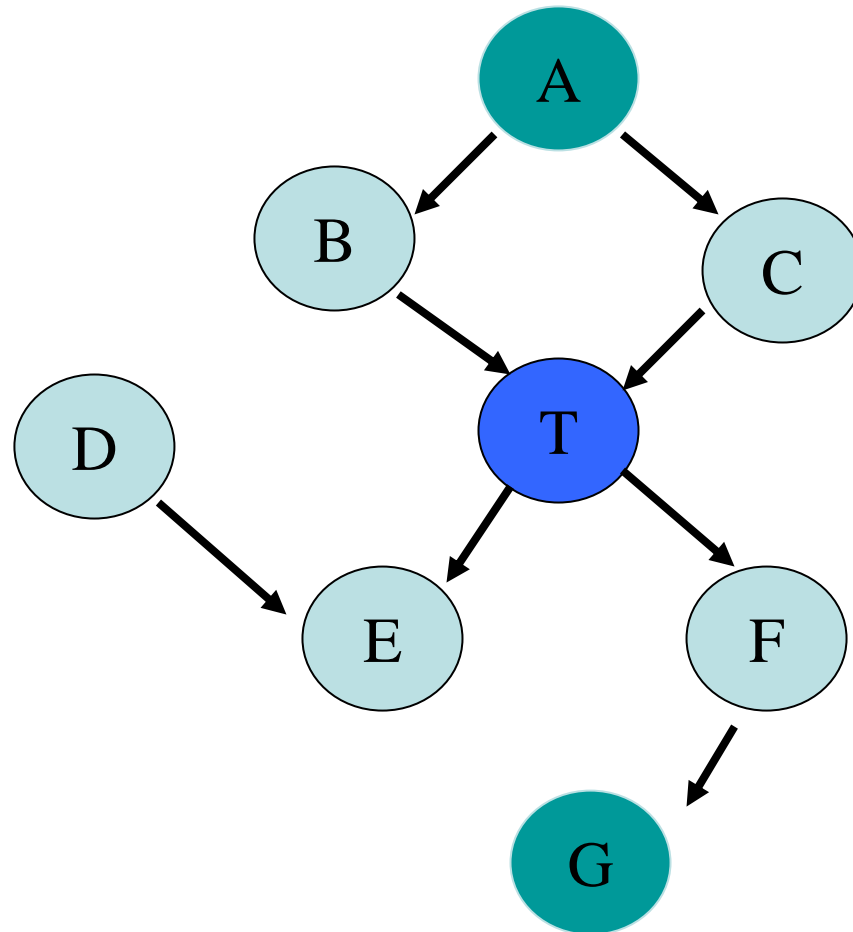
Variants:

1. Interleave phase I and II
2. Instead of II process with PC
3. Interleave I and PC

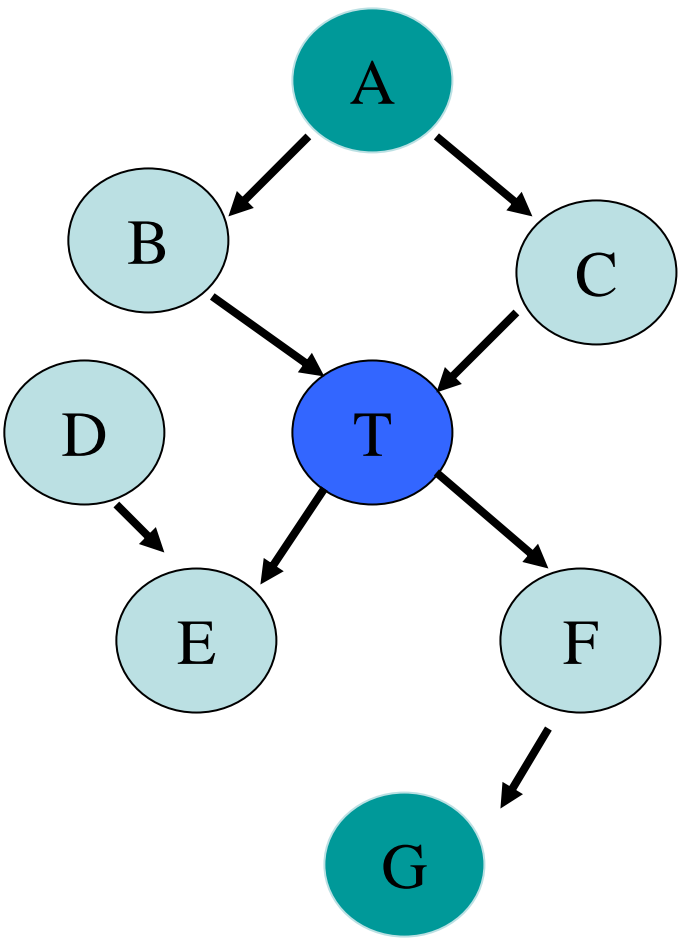
Example Trace of IAMB:

True structure depicted; members of the Markov Blanket of T are cyan

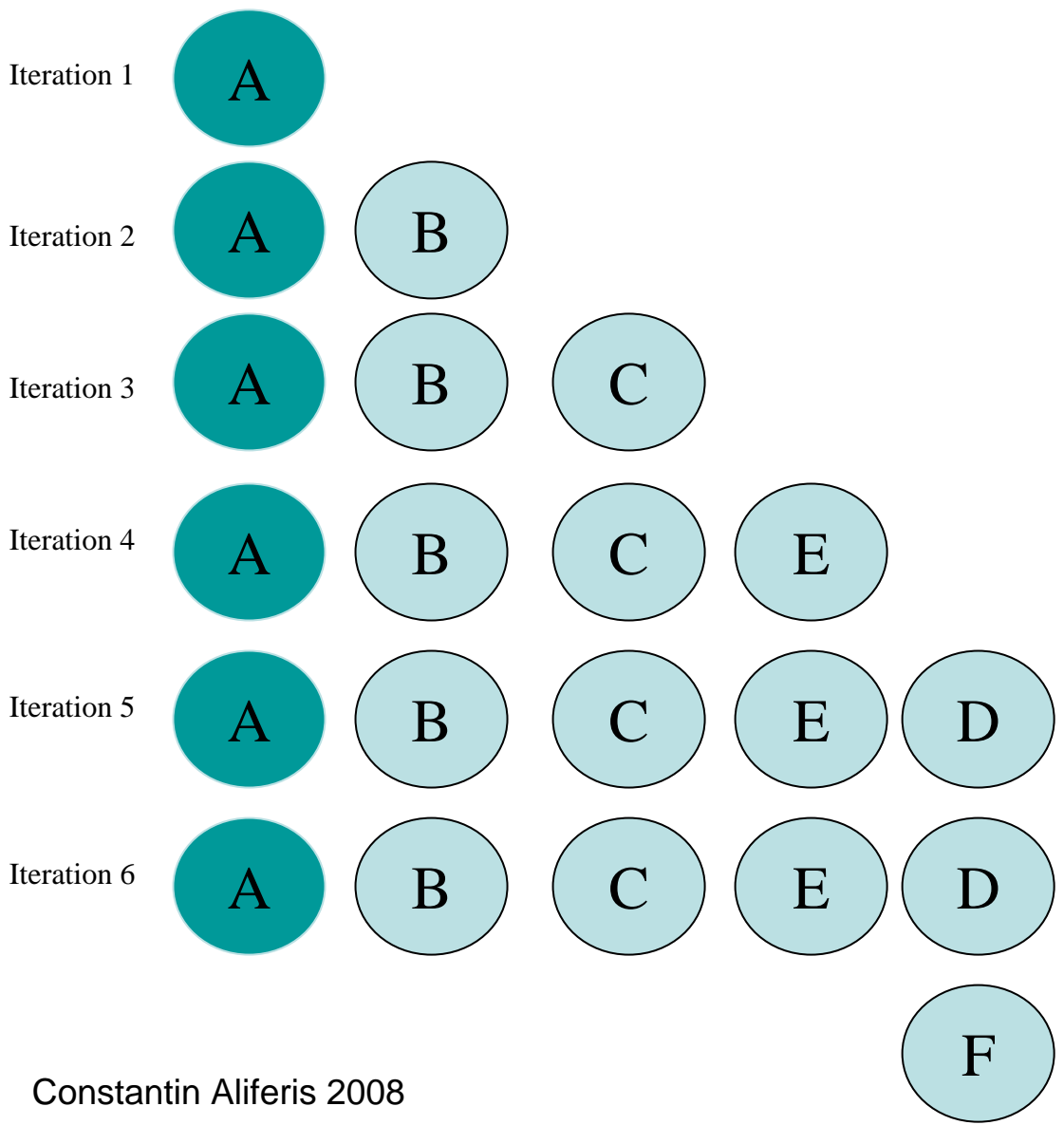
We have access to training data but not the true structure



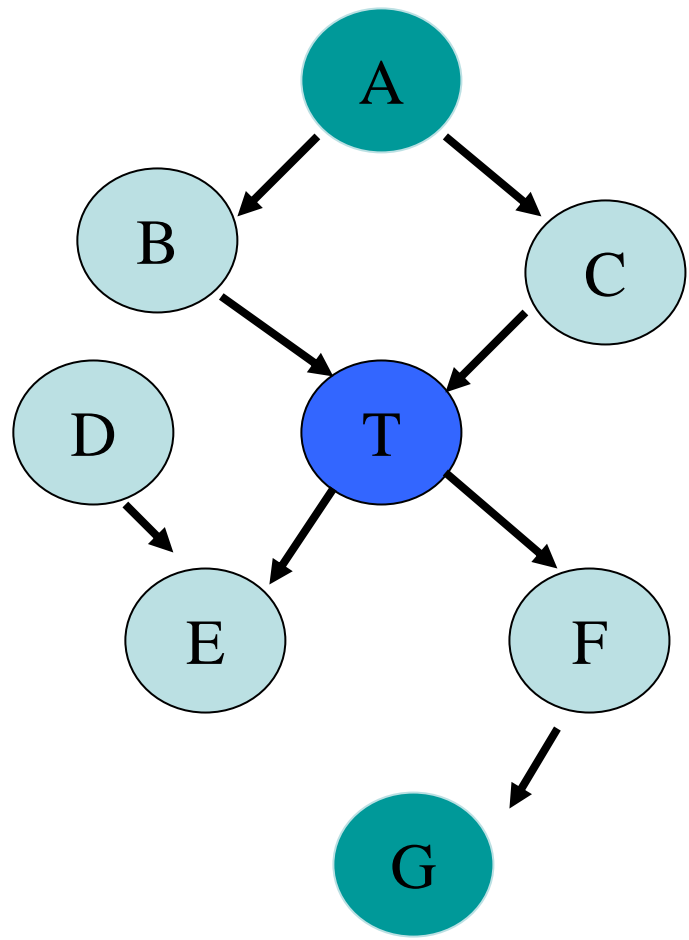
1. Phase I: build TMB



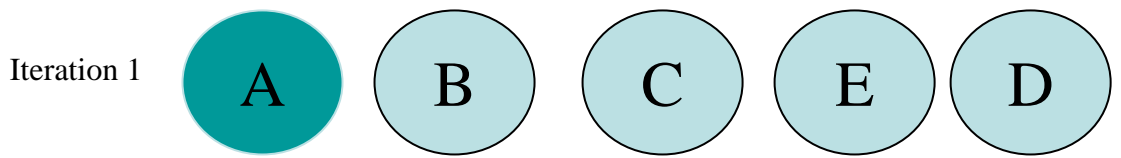
Tentative MB:



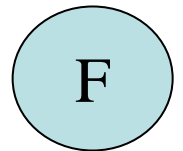
1. Phase II: remove false positives



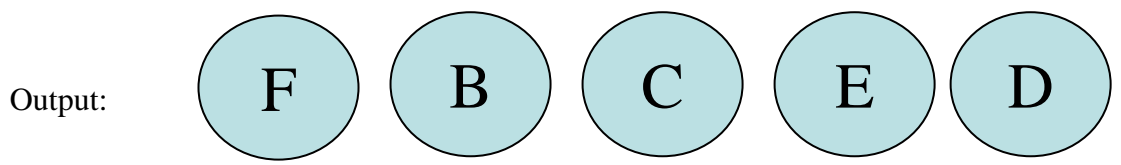
Tentative MB:



A is removed because
 $\perp(A : T \mid B, C, E, D, F)$



Iteration 2-6 No other variable can be removed



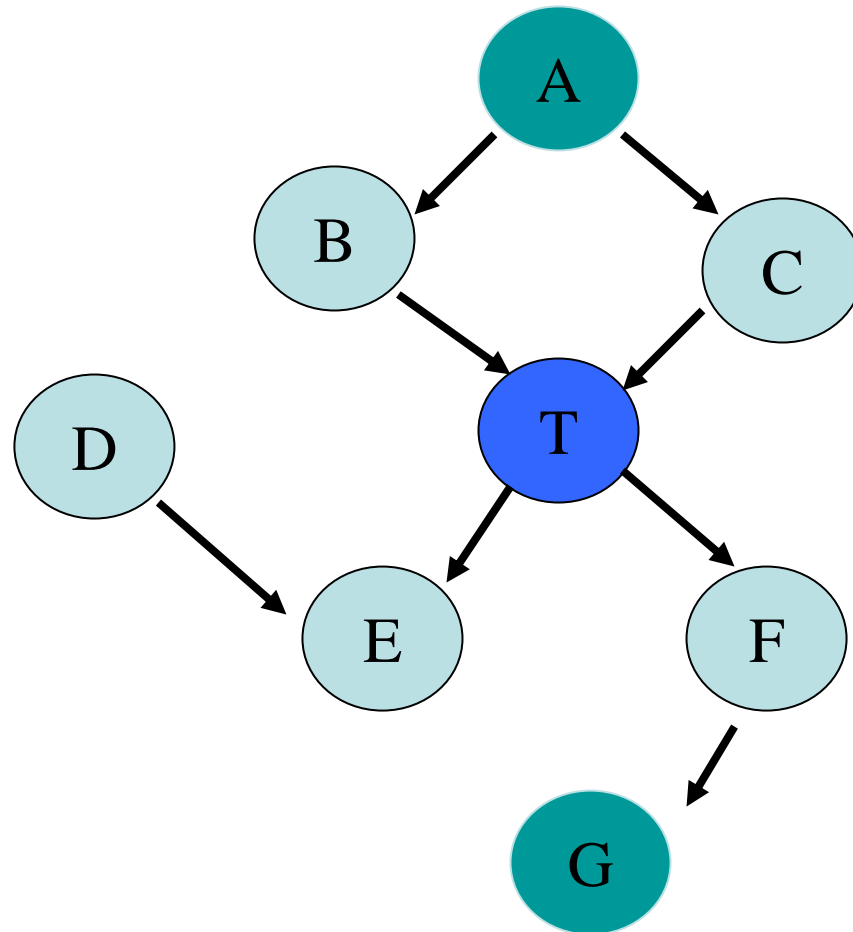
HITON-MB: when sample is small relative to $|MB(T)|$

(barring speed-up optimizations)

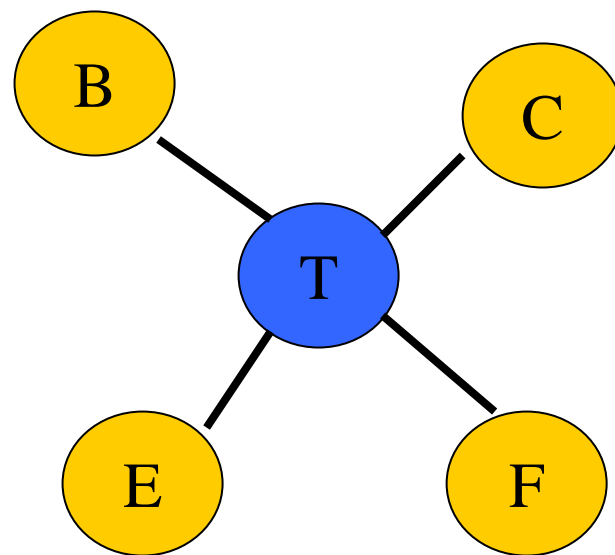
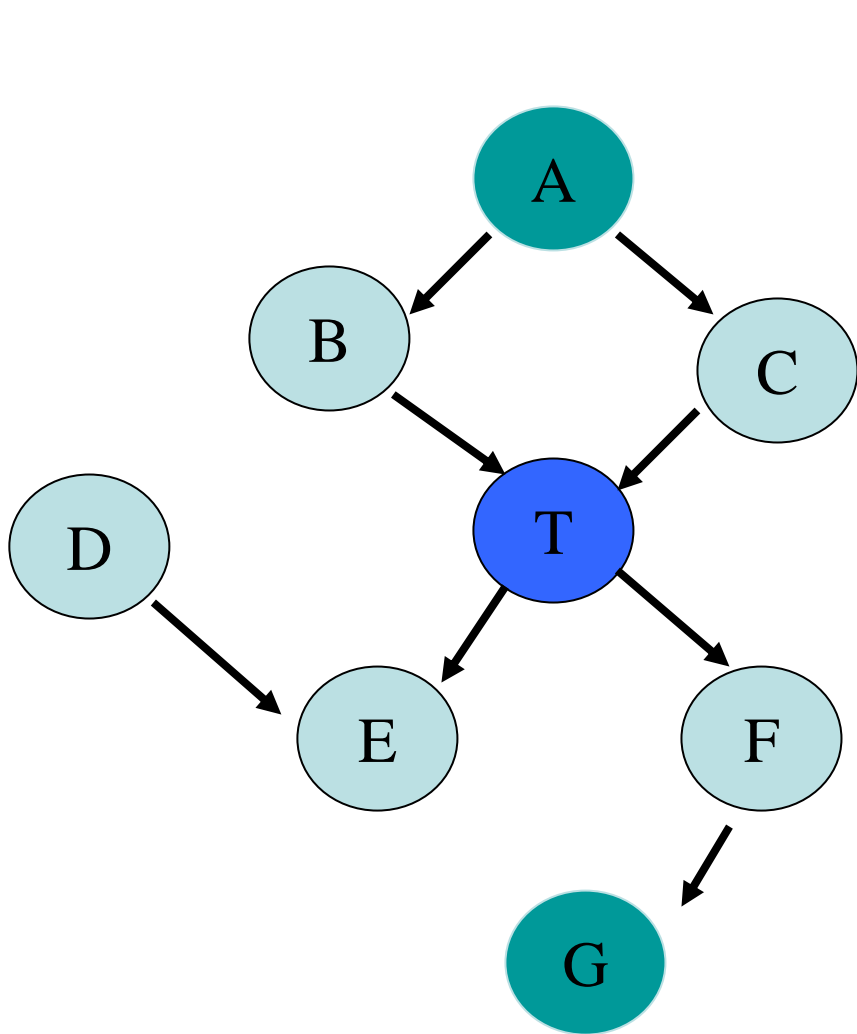
Example Trace of HITON:

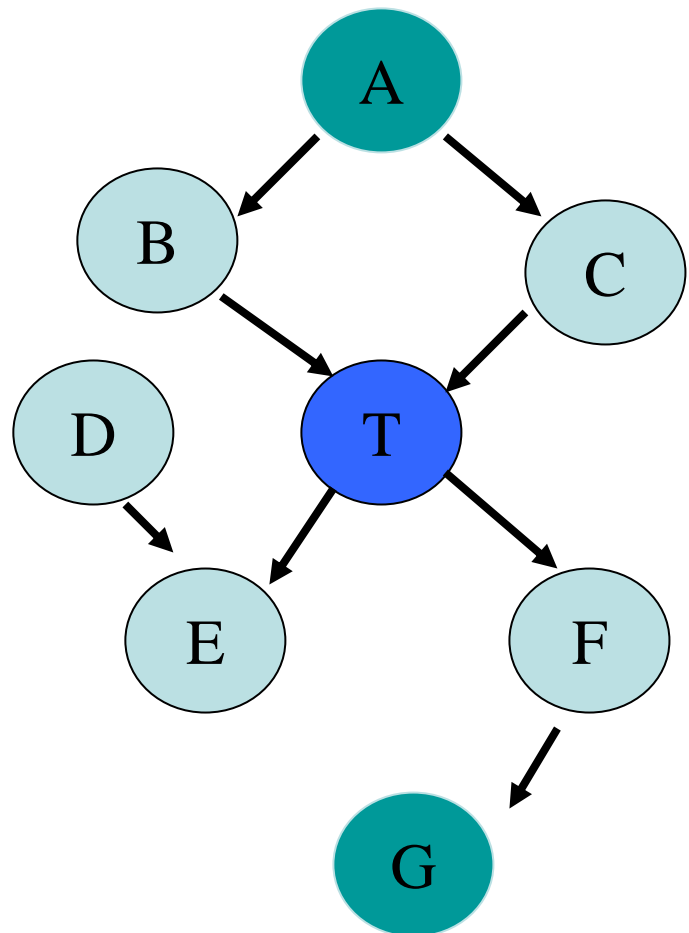
True structure depicted; members of the Markov Blanket of T are cyan

We have access to training data but not the true structure

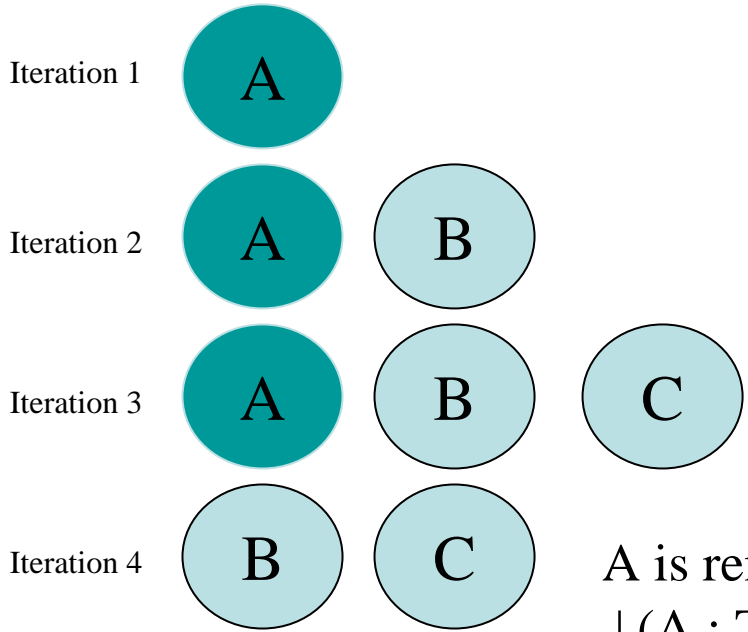


1. Identify variables with direct edges to the target T

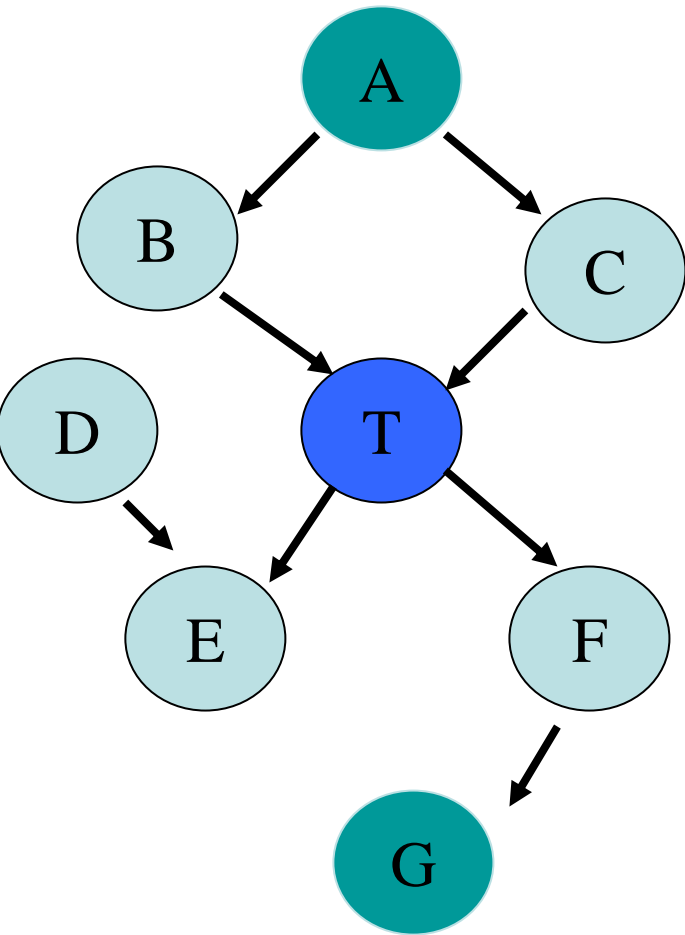




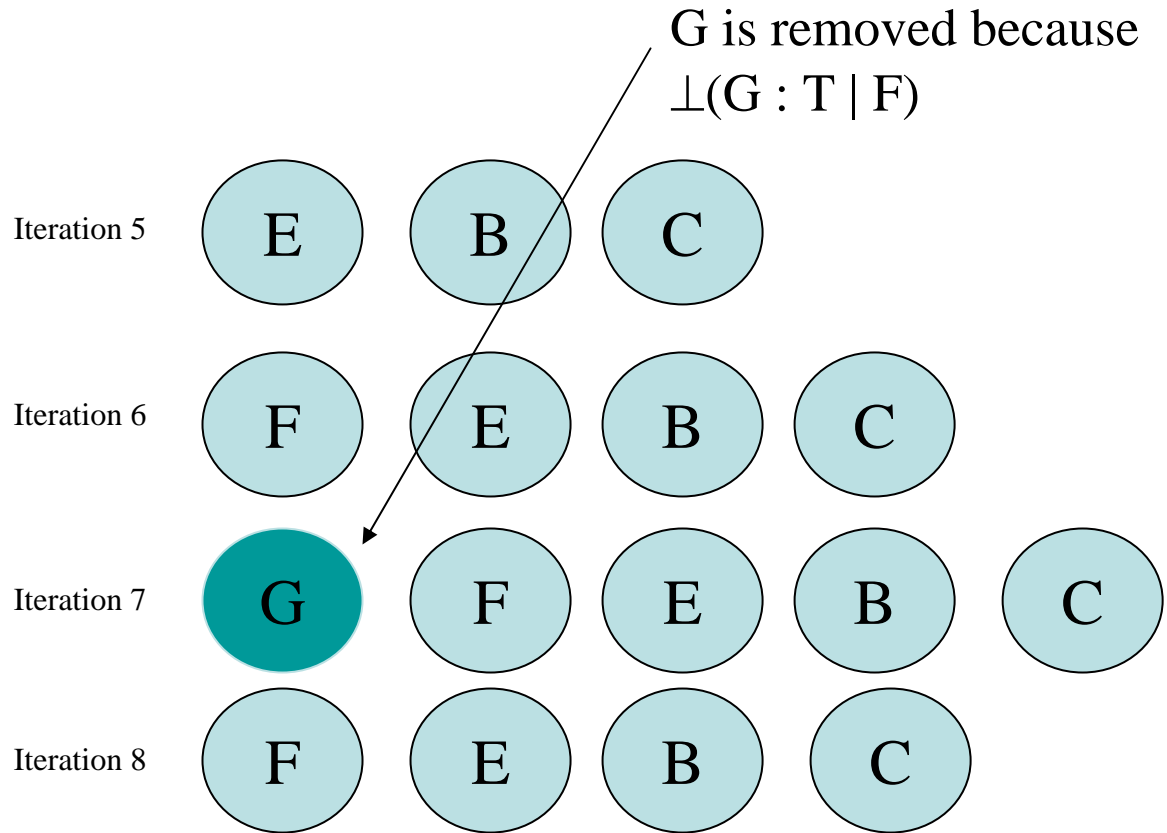
Tentative PC:



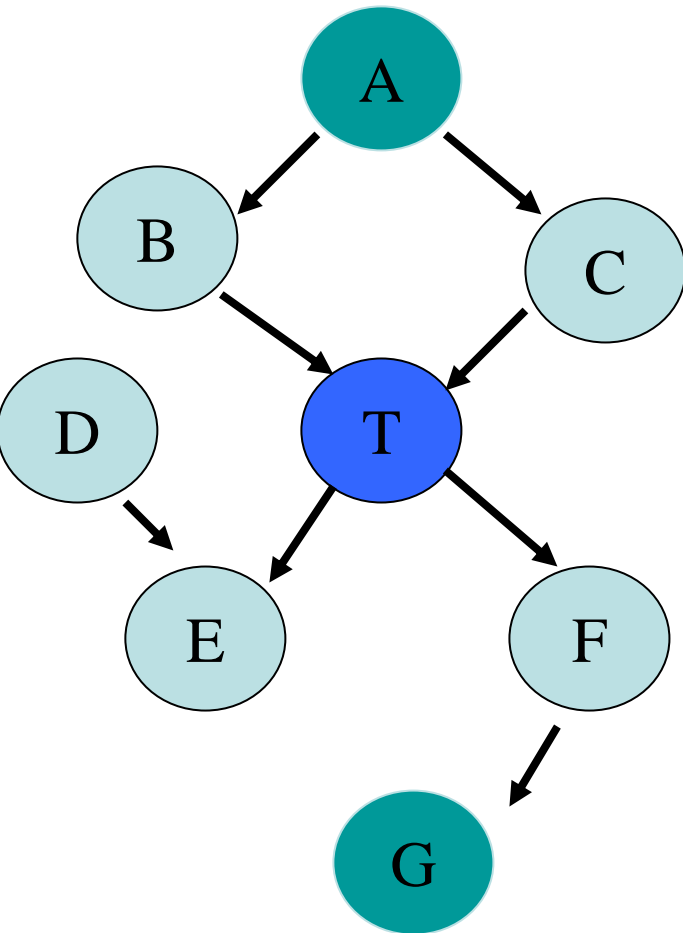
A is removed because $\perp(A : T \mid B, C)$



Tentative PC (continued):



Algorithm terminates because there are not other variables left to consider.



Symmetry:

When running the previous procedure for B returns: A, T.

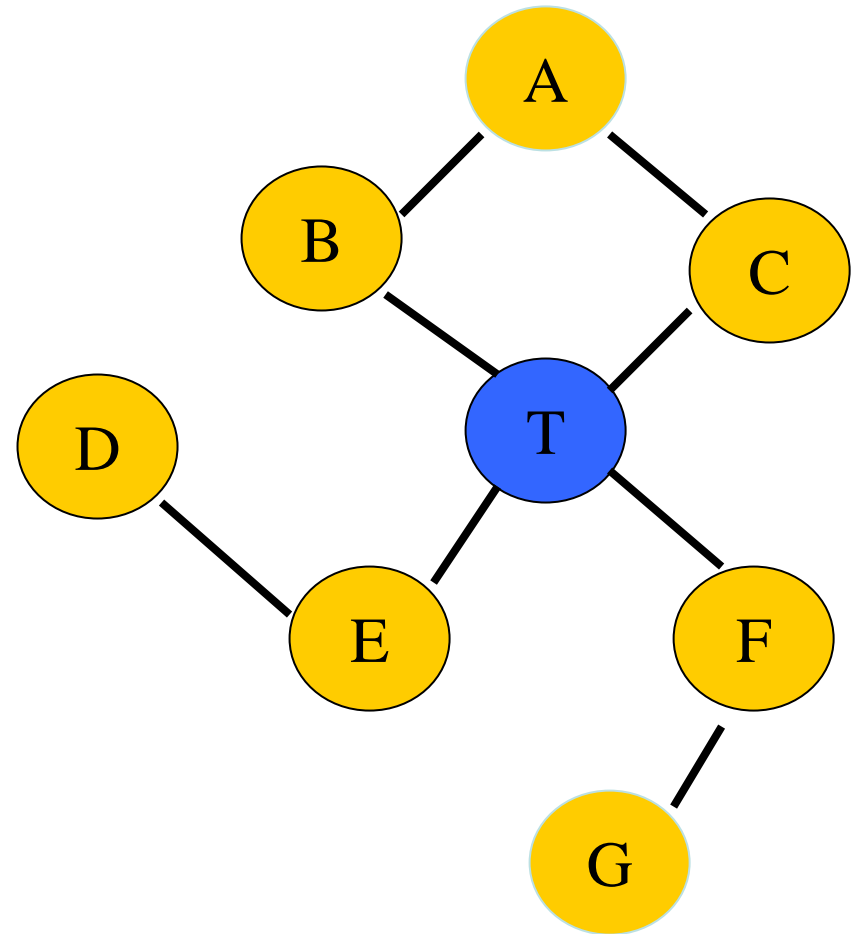
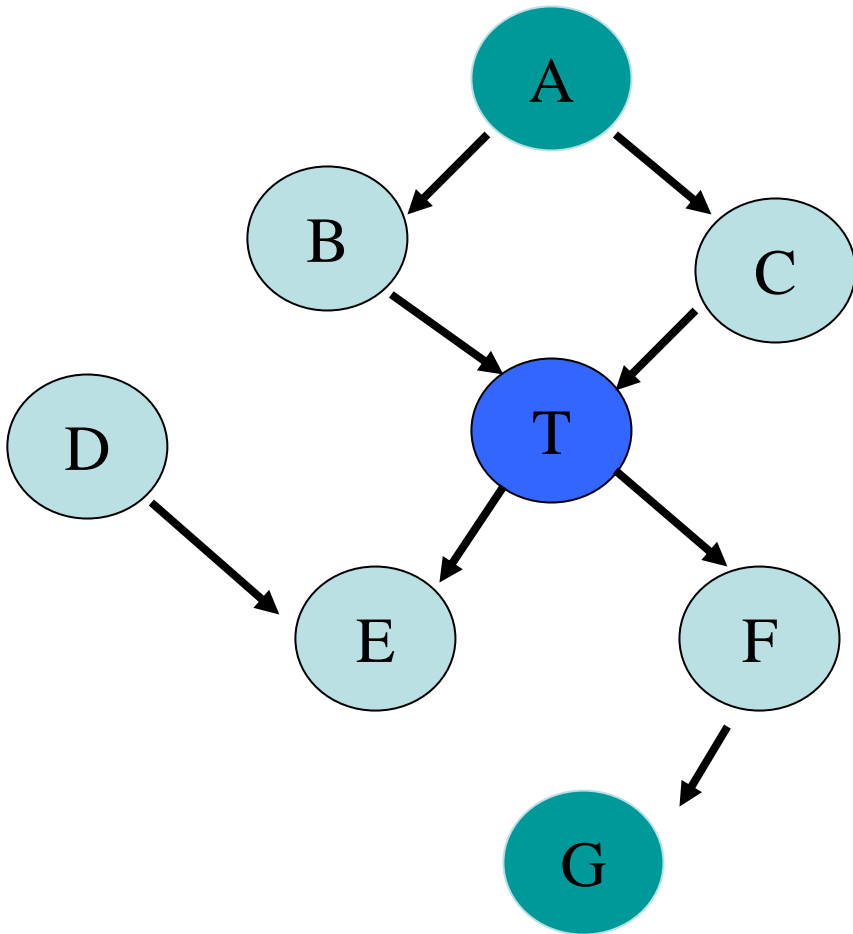
When running the previous procedure for C returns: A, T

When running the previous procedure for E returns: D, T.

When running the previous procedure for F returns: G, T.

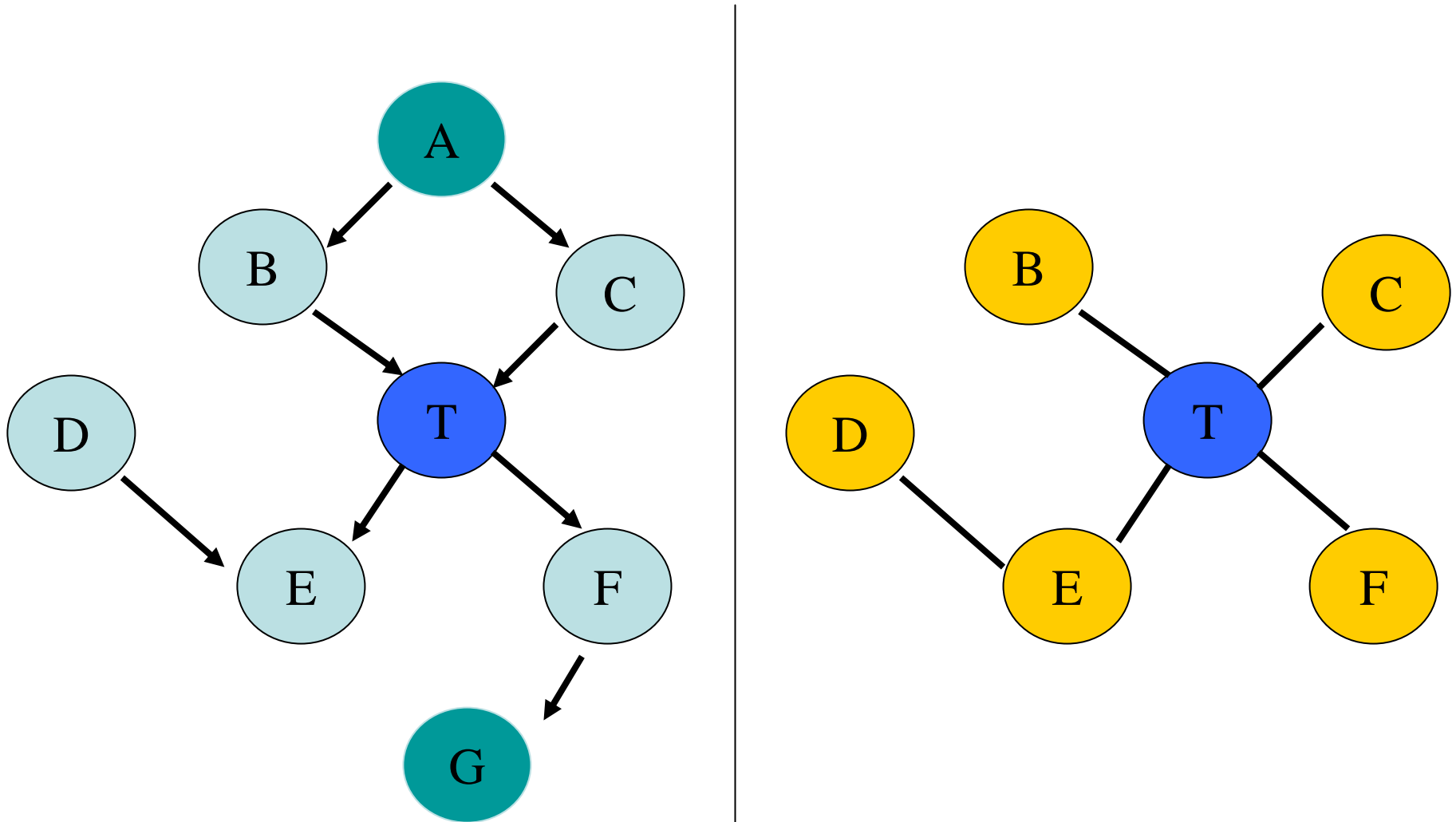
Hence all B,C,E,F satisfy symmetry and are retained.

2. Repeat previous for all members of PC and take the union of the resulting variables to be U.

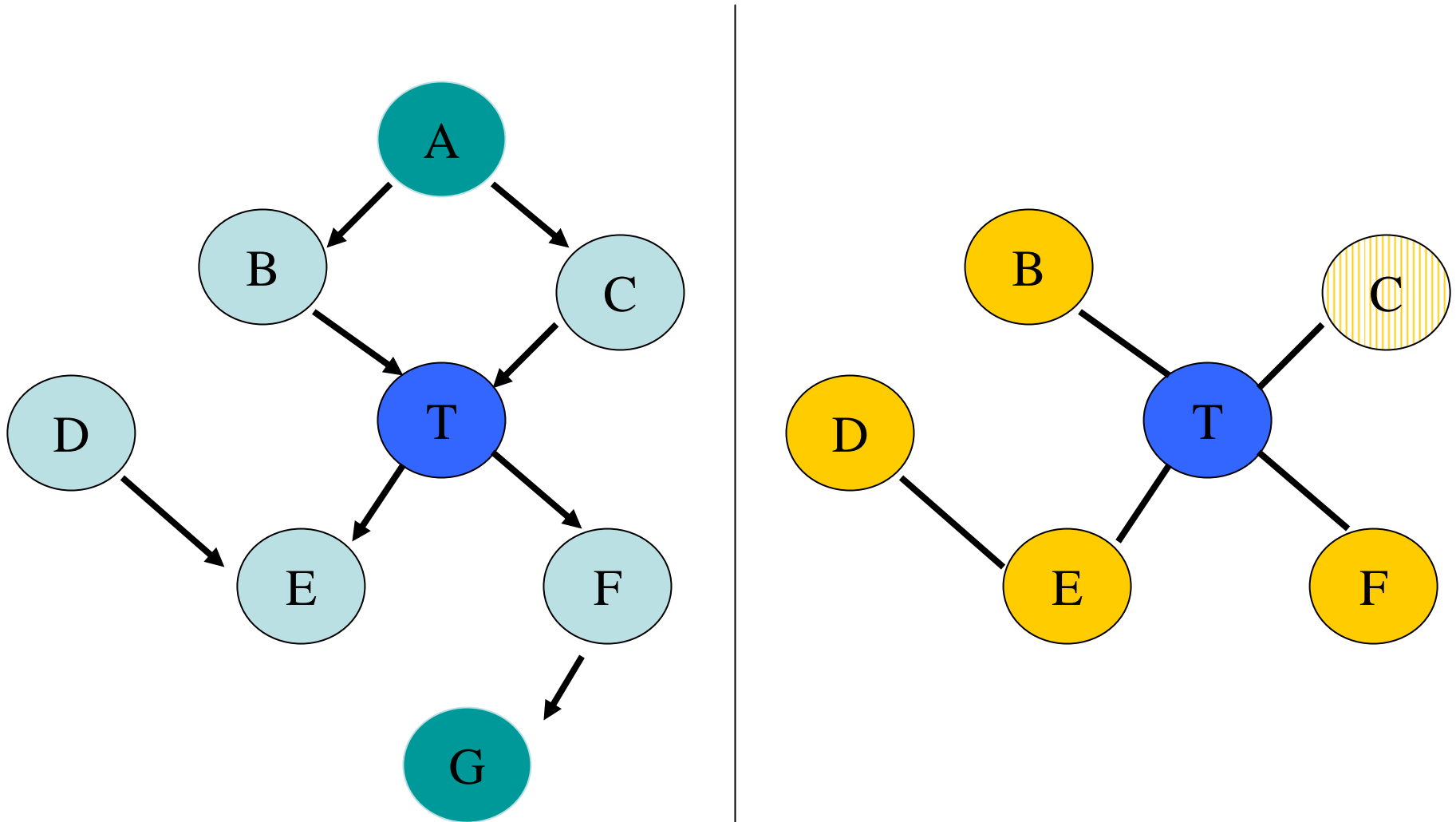


3. Throw away non-members of the Markov Blanket.

SGS criterion: A member X of PCPC that is not in PC is a member of the Markov Blanket if there is some member of PC Y , such that X becomes conditionally dependent with T conditioned on any subset of the remaining variables and Y .



4. Optional: If we desire to use the Markov Blanket for classification, eliminate any unnecessary variables by using a wrapping approach and cross-validation.



HITON-MB is just one
of infinite possibilities in its class

GLL-PC: Generalized Local Learning - Parents and Children

GLL-PC: High-level pseudocode and main components of Generalized Local Learning - Parents and Children. Returns $PC(T)$

- 1 $U \leftarrow \text{GLL-PC-nonsym}(T)$ // first approximate $PC(T)$ without symmetry check
- 2 For all $X \in U$
- 3 If $T \notin \text{GLL-PC-nonsym}(X)$ then $U \leftarrow U \setminus \{X\}$ // check for symmetry
- 4 Return U // true set of parents and children

GLL-PC-nonsym(T) // returns a set which is a subset of $EPC(T)$ and a superset of $PC(T)$

1. Initialization

- a. Initialize a set of candidates for the true $PC(T)$ set: $TPC(T) \leftarrow S$, s.t. $S \subseteq V\{T\}$
- b. Initialize a priority queue of variables to be examined for inclusion in $TPC(T)$: $OPEN \leftarrow V\{T \cup TPC(T)\}$

2. Apply inclusion heuristic function

- a. Prioritize variables in OPEN for inclusion in $TPC(T)$;
- b. Throw away non-eligible variables from OPEN;
- c. Insert in $TPC(T)$ the highest-priority variable(s) in OPEN and remove them from OPEN

3. Apply elimination strategy to remove variables from $TPC(T)$

4. Apply interleaving strategy by repeating steps #2 and #3 until a termination criterion is met

5. Return $TPC(T)$

Steps #1,2,3,4 can be instantiated in infinite ways.

There are rules that determine the admissible instantiations (which are themselves infinite)

GLL-PC: Admissibility rules

GLL-PC: Admissibility rules

1. The inclusion heuristic function should respect the following requirement:

// Admissibility rule #1

All variables $X \in PC(T)$ are eligible for inclusion in the candidate set $TPC(T)$ and each one is assigned a non-zero value by the ranking function. Variables with zero values are discarded and never considered again.

Note that variables may be re-ranked after each update of the candidate set, or the original ranking may be used throughout the algorithm's operation.

2. The elimination strategy should satisfy the following requirement:

// Admissibility rule #2

All and only variables that become independent of the target variable T given any subset of the candidate set $TPC(T)$ are discarded and never considered again (whether they are inside or outside $TPC(T)$).

3. The interleaving strategy iterates inclusion and elimination any number of times provided that iterating stops when the following criterion is satisfied:

//Admissibility rule #3

At termination no variable outside the set $TPC(T)$ is eligible for inclusion and no variable in the candidate set can be removed at termination.

Respecting the admissibility rules of GLL-PC

- Obtain correct local causal neighborhood (direct causes and direct effects) under the following sufficient conditions:
 - Faithful distributions,
 - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size)
 - Local causal sufficiency (i.e., no confounders among direct causes/effects and the target).

HITON-PC as instance of GLL-PC

Interleaved HITON-PC with symmetry correction

Derived from GLL-PC with following instantiation specifics:

Initialization

$$TPC(T) \leftarrow \emptyset$$

Inclusion heuristic function

- Sort in descending order the variables X in OPEN according to their pairwise association with T , i.e., $Assoc(X, T/\emptyset)$.
- Remove from OPEN variables with zero association with T , i.e., when $I(X, T/\emptyset)$
- Insert at end of $TPC(T)$ the first variable in OPEN and remove it from OPEN

Elimination strategy

For each $X \in TPC(T)$

If $\exists Z \subseteq TPC(T) \setminus \{X\}$, s.t. $I(X, T/Z)$ remove X from $TPC(T)$

Interleaving strategy

Repeat

steps #2 and #3 of GLL-PC-nonsym

Until $OPEN = \emptyset$

This we call: interleaved HITON-PC with symmetry correction and is a correct algorithm.

MMPC as instance of GLL-PC

MMPC with symmetry correction

Derived from GLL-PC with following instantiation specifics:

Initialization

$TPC(T) \leftarrow \emptyset$

Inclusion heuristic function

- Sort in descending order the variables X in OPEN according to $\text{Min}_Z \text{Assoc}(X, T/Z)$ for $Z \subseteq TPC(T) \setminus \{X\}$
- Remove from OPEN variables with zero association with T , given some $Z \subseteq TPC(T) \setminus \{X\}$
- Insert at end of $TPC(T)$ the first variable in OPEN and remove it from OPEN

Elimination strategy

If OPEN = \emptyset

For each $X \in TPC(T)$

If $\exists Z \subseteq TPC(T) \setminus \{X\}$, s.t. $I(X, T/Z)$ remove X from $TPC(T)$

Interleaving strategy

Repeat

steps #2 and #3 of GLL-PC-nonsym

Until OPEN = \emptyset

This we call: MMPC with symmetry correction and is a correct algorithm.

GLL-MB: Generalized Local Learning – Markov Blanket

GLL-MB: Generalized Local Learning - Markov Blanket

1. $PC(T) \leftarrow GLL-PC(T)$ // obtain $PC(T)$ by running $GLL-PC$ for variable T
2. For every variable $Y \in PC(T)$
 $PC(Y) \leftarrow GLL-PC(Y)$ // obtain $PC(Y)$ for every member Y of $PC(T)$
3. $TMB(T) \leftarrow PC(T)$ // initialize $TMB(T)$ with $PC(T)$ members
4. $S \leftarrow \{\cup_{Y \in PC(T)} PC(Y)\} \setminus \{PC(T) \cup \{T\}\}$ // these are the potential spouses
5. For every variable $X \in S$
 - a. Retrieve a subset Z s.t. $I(X, T / Z)$ // subset was identified and stored in steps 1 and 2
 - b. For every variable $Y \in PC(T)$ s.t. $X \in PC(Y)$ // Y is a potential common child of T and X
 - c. If $\neg I(X, T | Z \cup \{Y\})$ // X is a spouse
 - d. Insert X into $TMB(T)$
6. Optionally: Eliminate from $TMB(T)$ predictively redundant members using a backward wrapper approach.
7. Return $TMB(T)$

Steps #1,6 can be instantiated in infinite ways.

Admissibility requirements: use an admissible GLL-PC and a sufficiently powerful wrapper.

Respecting the admissibility rules of GLL-MB

- Obtain correct minimal Markov Blanket under the following sufficient conditions :
 - Faithful distributions,
 - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size).

HITON-MB as instance of GLL-MB

Derived from GLL-MB with following instantiation specifics:

GLL-PC is instantiated as interleaved HITON-PC with symmetry.

This we call: interleaved HITON-MB with symmetry correction and is a correct algorithm.

LGL: Locally-constrained Global Learning

LGL: Local-to-Global Learning

1. Find $PC(X)$ for every variable X in the data using an admissible instantiation of GLL-PC and prioritizing which variables to induce $PC(X)$ for, according to a prioritization strategy.
2. Piece together the undirected skeleton from the local GLL-PC results.
3. Use any desired arc orientation scheme to orient edges.

#1,2,3 can be instantiated in infinite ways. If an admissible GLL-PC is used in #1, and admissible orientation scheme in #3, then the total algorithm is admissible.

Respecting the admissibility rules of LGL

- Obtain correct causal graph under the following sufficient conditions :
 - Faithful distributions,
 - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size); alternatively correct statistical decisions about graph structure scoring.
 - Causal sufficiency (i.e., no confounders between any pair of variables).

MMHC: instance of LGL

MMHC Global Learning Algorithm

1. Find $PC(X)$ for every variable X in data using MMPC (without symmetry correction) and lexicographic prioritization.
2. Piece together the undirected skeleton using an “OR rule” (an edge exists between A and B iff A is in $PC(B)$ or B is in $PC(A)$).
3. Use greedy steepest-ascent TABU search and BDeu score to orient edges.

MMHC is inadmissible with respect to both the skeleton and with respect to orientation.

Presentations: Experimental results & other details