

Markov Blanket Methods for Classification: Applications in the Molecular Diagnosis of Lung Cancer and Thrombin Binding

C.F. Aliferis M.D., Ph.D.

October 16, 2002

Collaborators:

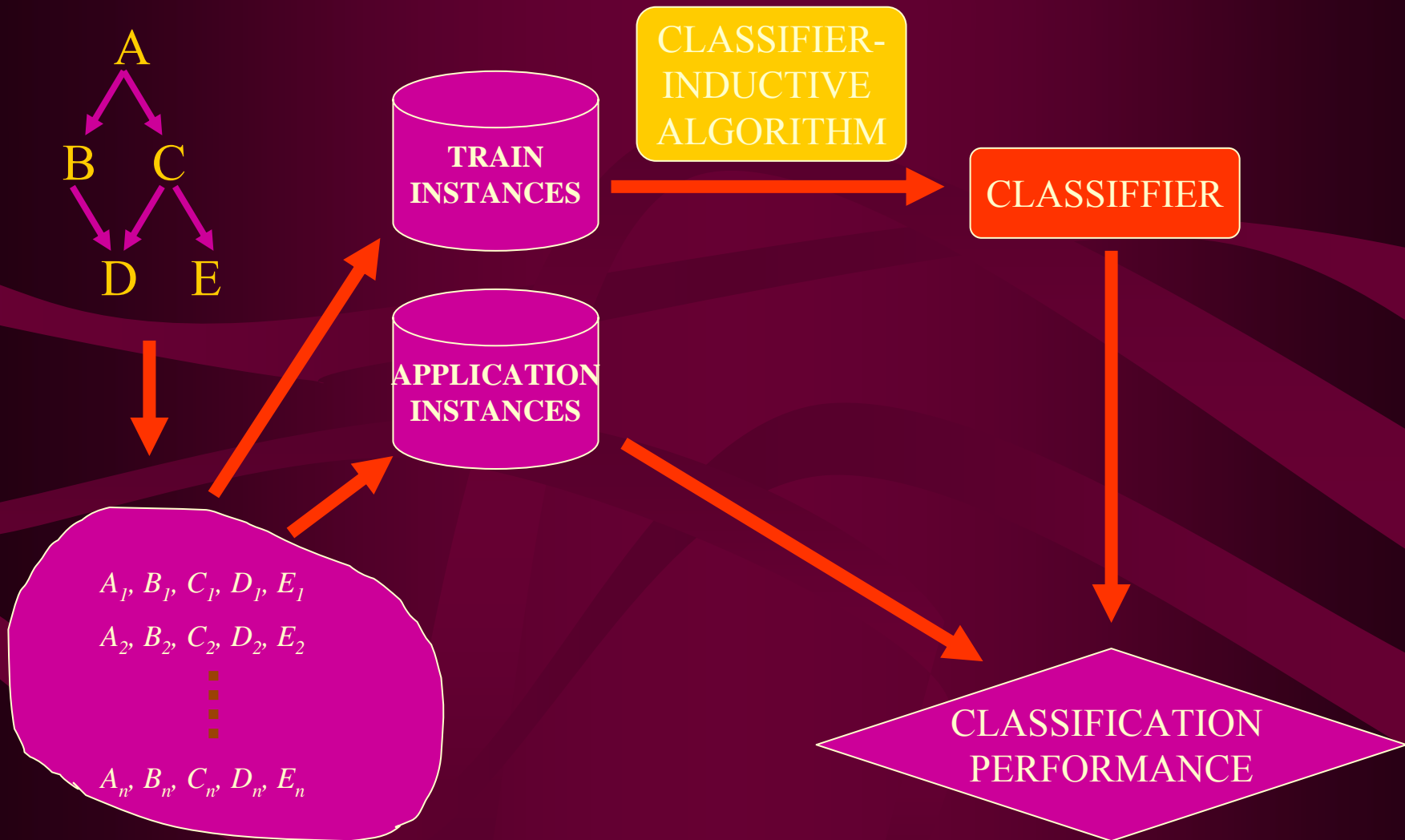
- I. Tsamardinos Ph.D.,
- Pierre Massion M.D.,
- Alexander Statnikov M.S.,
- Nafeh Fananapazir

Goals are to show that:

1. By using novel Markov Blanket Methods we can derive optimal prediction and diagnostic models in gene expression and structural biology tasks
2. The Markov Blanket methods achieve excellent to outstanding reduction in the number of predictors needed
3. The Markov Blanket methods are efficient in time and sample even in the presence of massive numbers of variables
4. The Markov Blanket methods offer additional information than state-of-the art gene selection methods
5. The biological interpretation is clearer than existing methods since, contrary to the latter, it is tied to causal structure

Informatics: Background & Methods

Classification



What Good Is A Classifier For?

- Diagnosis
- Prognosis
- Treatment Selection
- Predict Response to Treatment

What is Feature Selection for classification?

- Given: a set of predictors (“features”) V and a target variable T
- Find: minimum set F that achieves maximum classification performance of T

Why feature selection is important?

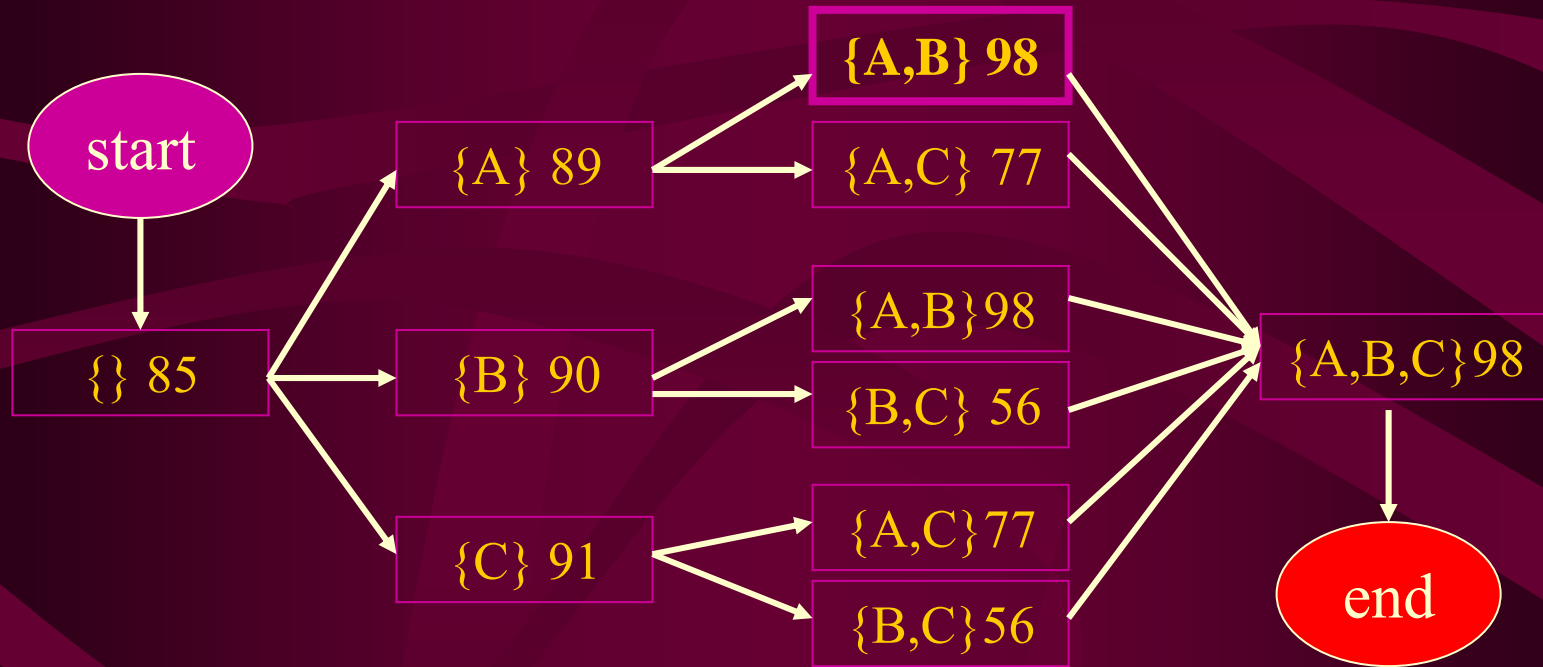
- May Improve performance of classification algorithm
- Classification algorithm may not scale up to the size of the full feature set either in available sample or time
- Allows us to better understand the domain
- Cheaper to collect a reduced set of predictors
- Safer to collect a reduced set of predictors

Classic Approach to Feature selection: Wrappers

Suppose we have predictors A, B, C and classifier M . We want to predict T given the smallest possible subset of $\{A,B,C\}$, while achieving maximal performance

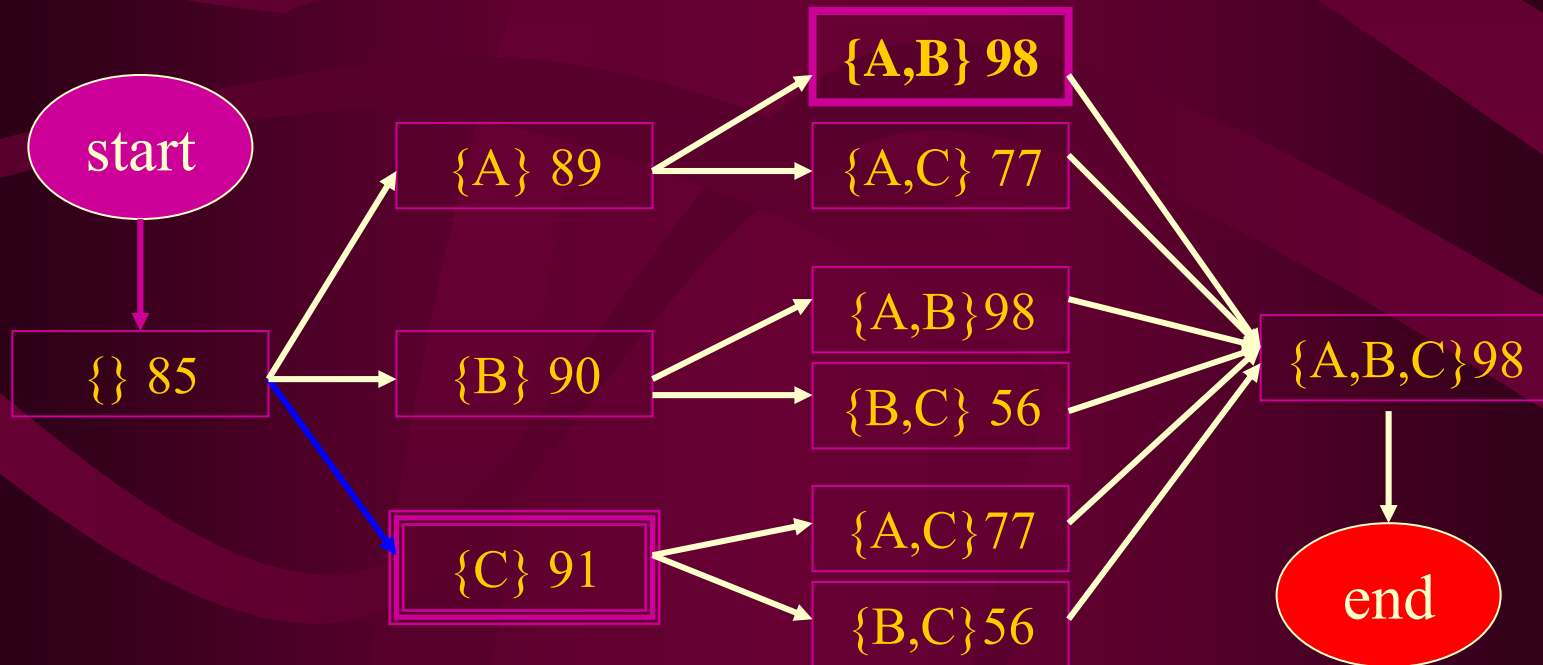
FEATURE SET	CLASSIFIER	PERFORMANCE
$\{A,B,C\}$	M	<u>98%</u>
<u>$\{A,B\}$</u>	M	<u>98%</u>
$\{A,C\}$	M	77%
$\{B,C\}$	M	56%
$\{A\}$	M	89%
$\{B\}$	M	90%
$\{C\}$	M	91%
$\{.\}$	M	85%

Classic Approach to Feature selection: Wrappers



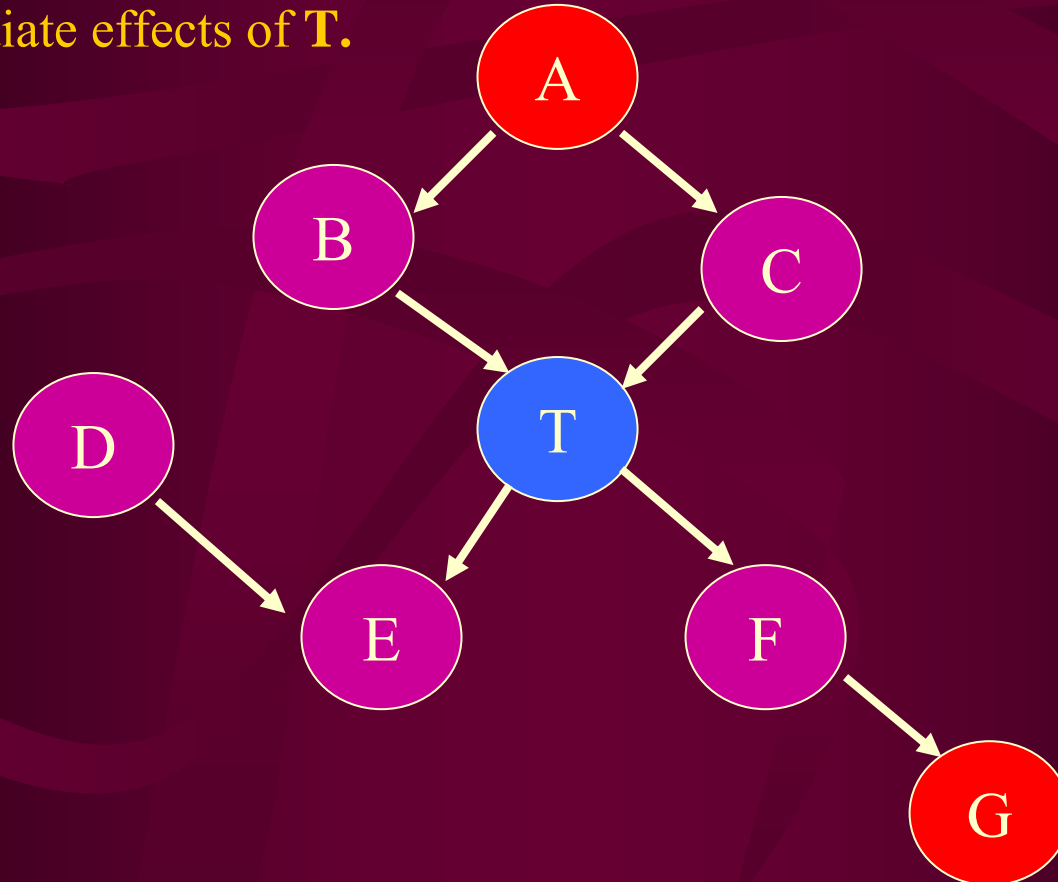
Classic Approach to Feature selection: Wrappers

The set of all subsets is the power set and its size is $2^{|V|}$. Hence for large V we cannot do this procedure exhaustively; instead wrappers employ *heuristic search* of the space of all possible feature subsets. A common example of heuristic search is hill climbing: keep adding features one at a time until no further improvement can be achieved.

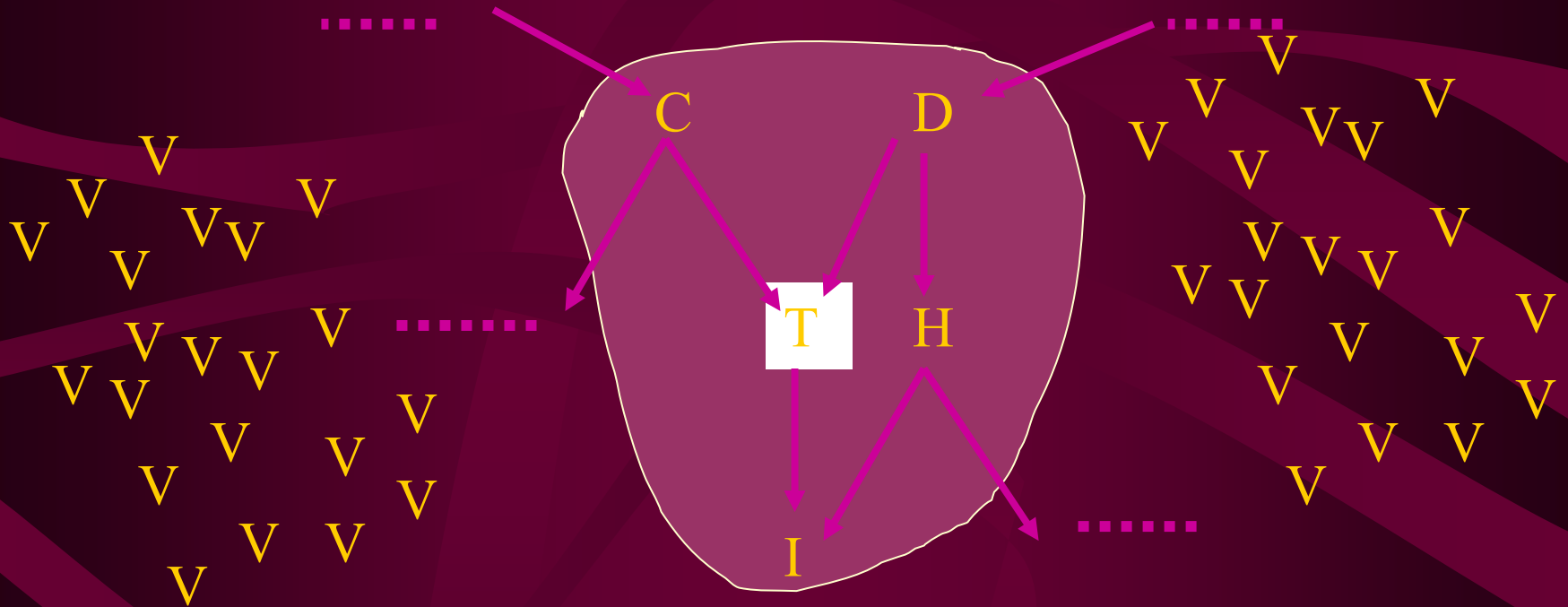


How do we approach the feature selection problem in Discovery Systems Lab research?

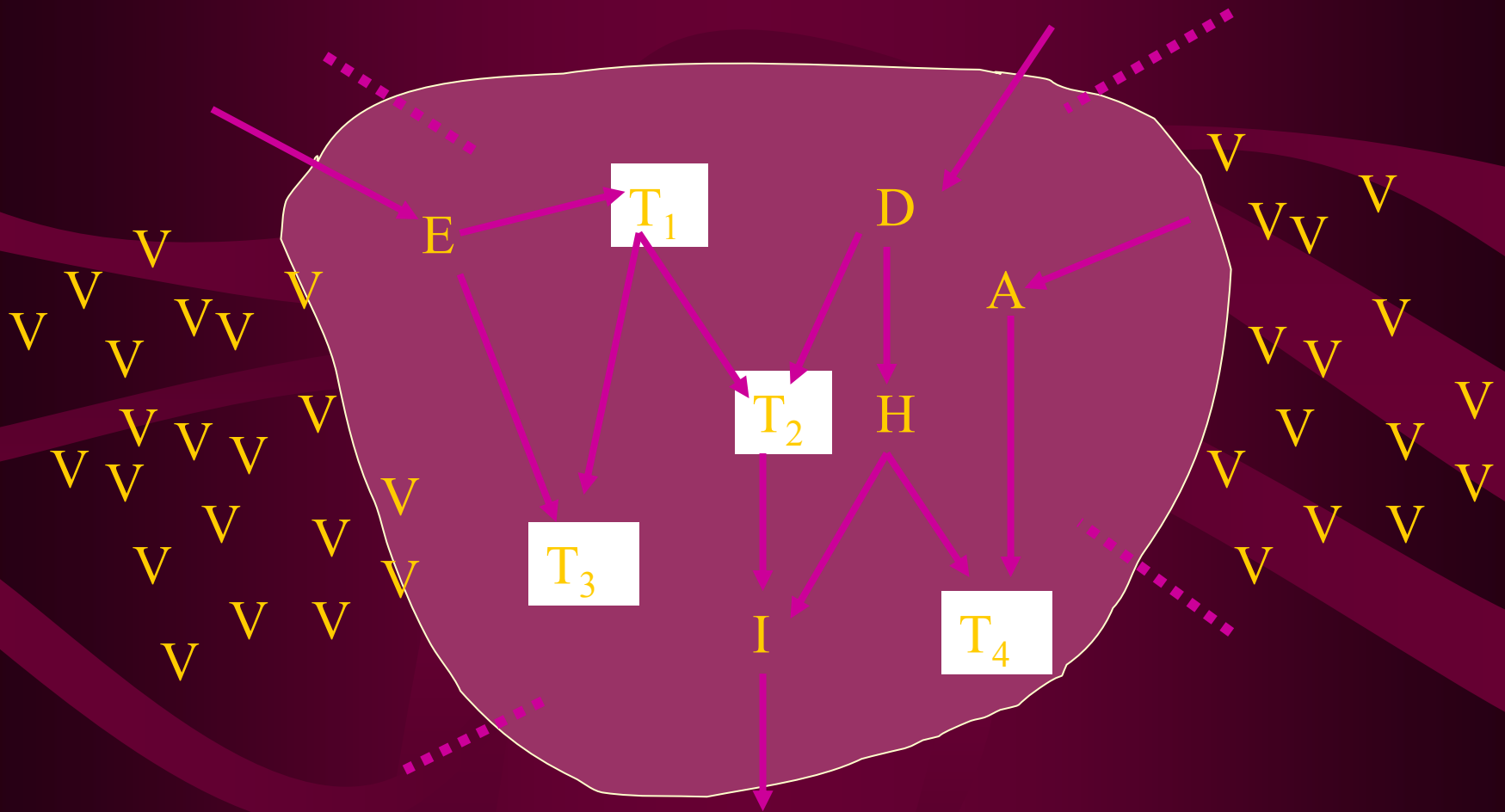
- (Reminder) Definition:
 - The Markov Blanket of some variable of interest T (“MB(T)”) is the set of the immediate causes, immediate effects, and immediate causes of the immediate effects of T .



Fundamental Concept: Predicting Individual Micro-States (variables, e.g., Genes)



Fundamental Concept: Predicting Macro-States (e.g., Disease)



A Crucial property of the Markov Blanket

$MB(T)$ is the minimal set of predictor variables needed for classification (diagnosis, prognosis, etc.) of the target variable T

So, One Way To Solve The Feature Selection Problem:

- Goal:
 - Given: Data (observations of T and a set of variables V)
 - Find: $MB(T)$

How Can One Find the Markov Blanket?

- Previously MB(T) could be discovered using a full-network induction algorithm, or various heuristic procedures
- The state-of-the-art (full-network) algorithms try to learn the whole network and are not tractable for large networks
- New algorithms developed in our lab induce the MB directly and are highly-scalable without compromising soundness

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)

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 $f(V_i ; T | \text{CurrentMB})$

// 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

If not $I(V_i ; T | \text{CurrentMB})$ Admit candidate variable V_i into *CurrentMB*,

Else Exit Loop

Until False

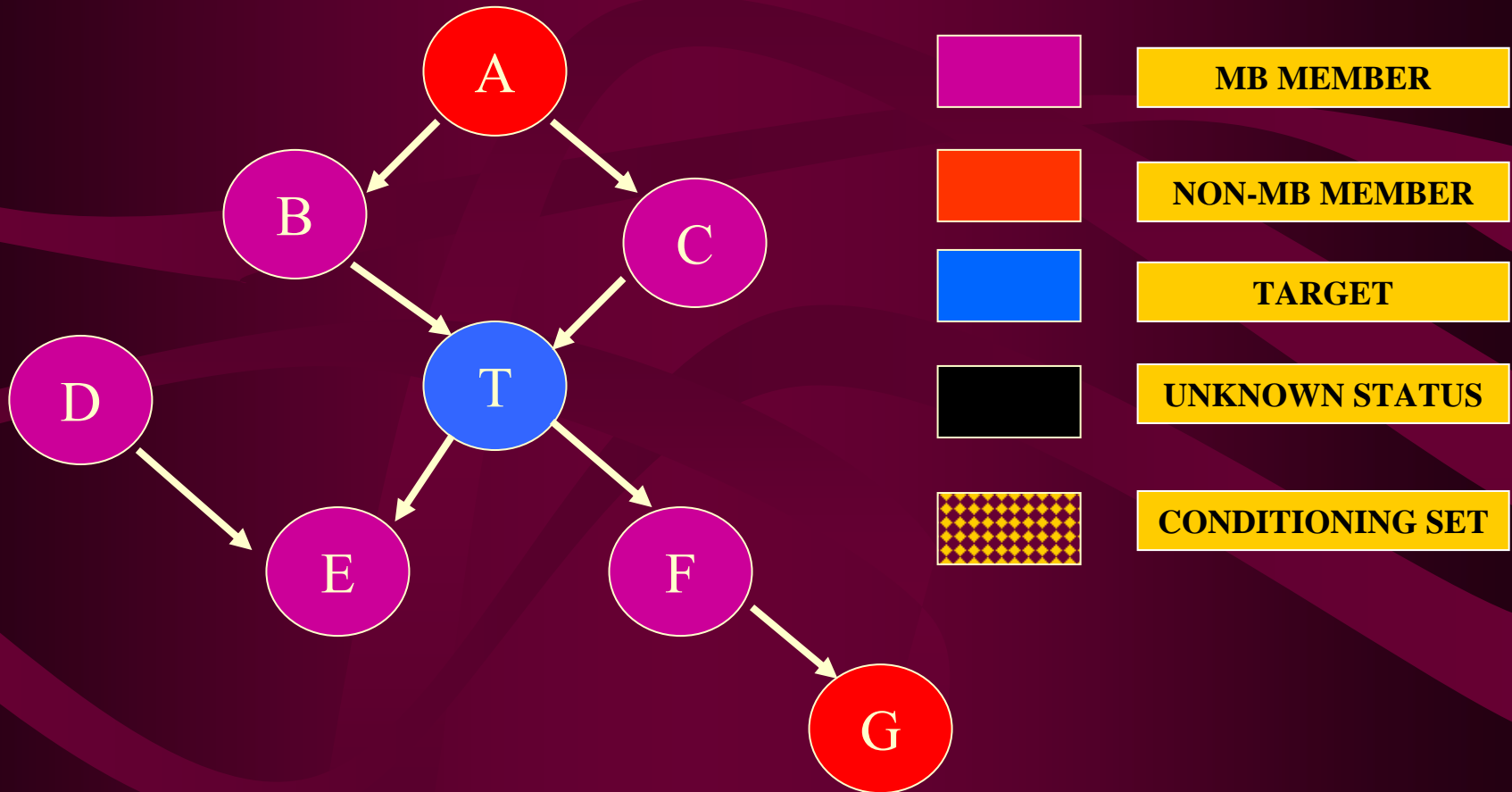
Phase II:

For all members V_j of *CurrentMB*

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

Return *CurrentMB*

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



Statistical Reminder: Conditional independence

- Two variables X and Y are conditionally independent given Z , denoted as $I(X; Y | Z)$, iff the probability distribution of X is the same for all values of Y and this holds for each value of Z :
- Intuitive Meaning: given that I know Z (“conditioned on Z ”), X does not give me information about Y (X is “uninformative”, “non-predictive”, “independent” of Y)

$$p(X+, Y+)$$

$$=$$

$$p(X+, Y-)$$

$$Z+$$

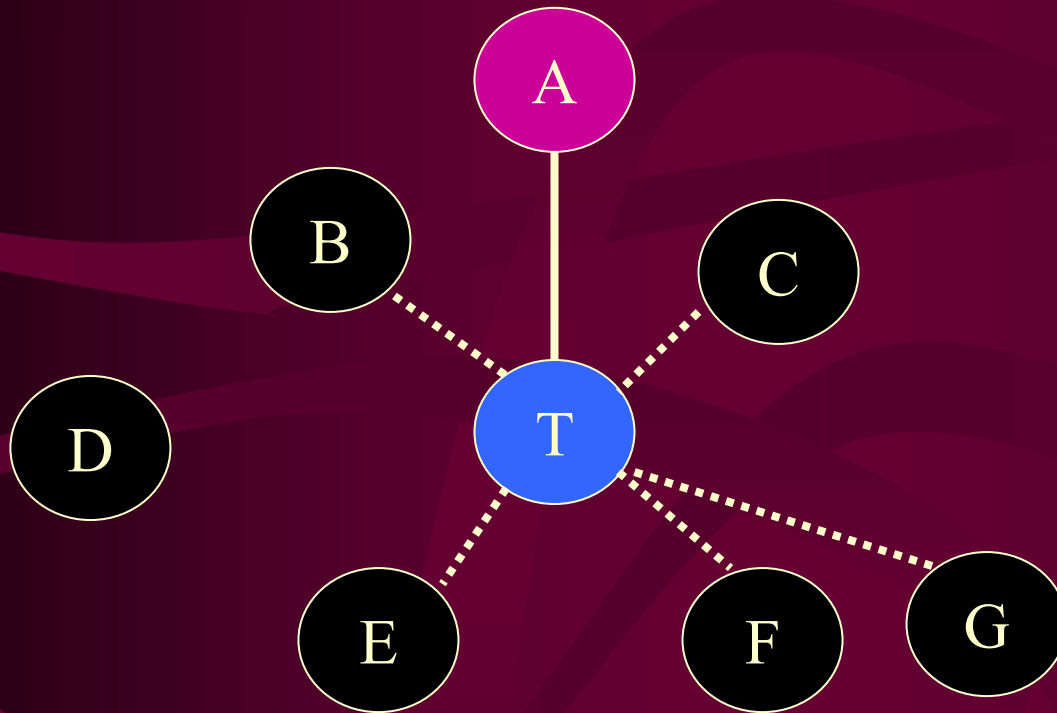
$$p'(X+, Y+)$$

$$=$$

$$p'(X+, Y-)$$

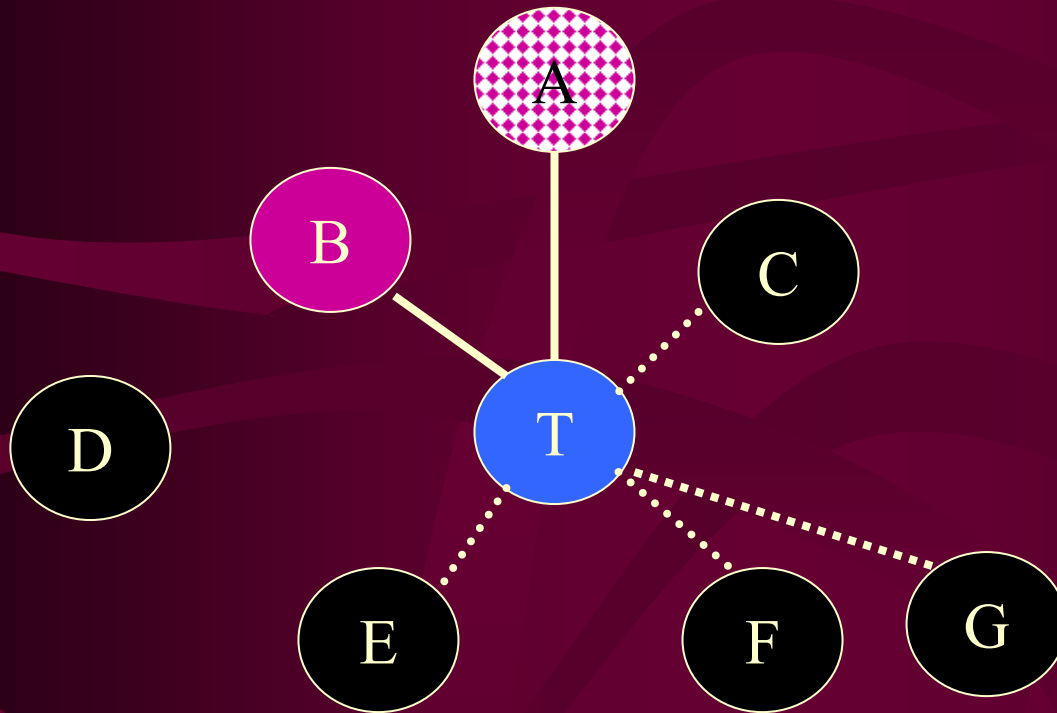
$$Z-$$

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



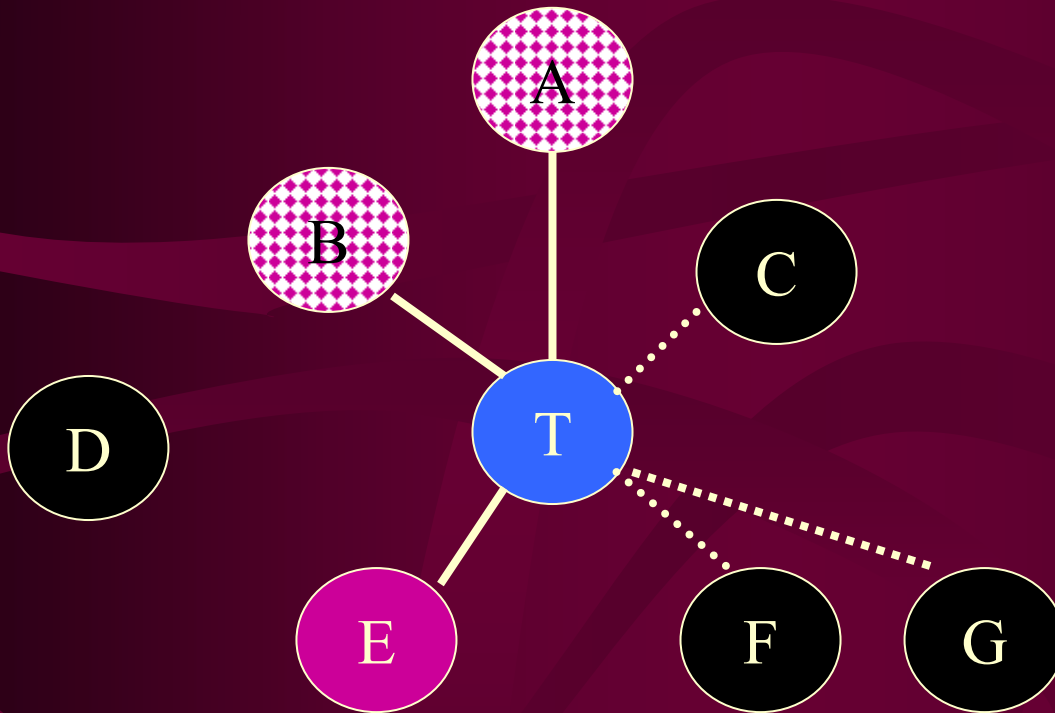
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



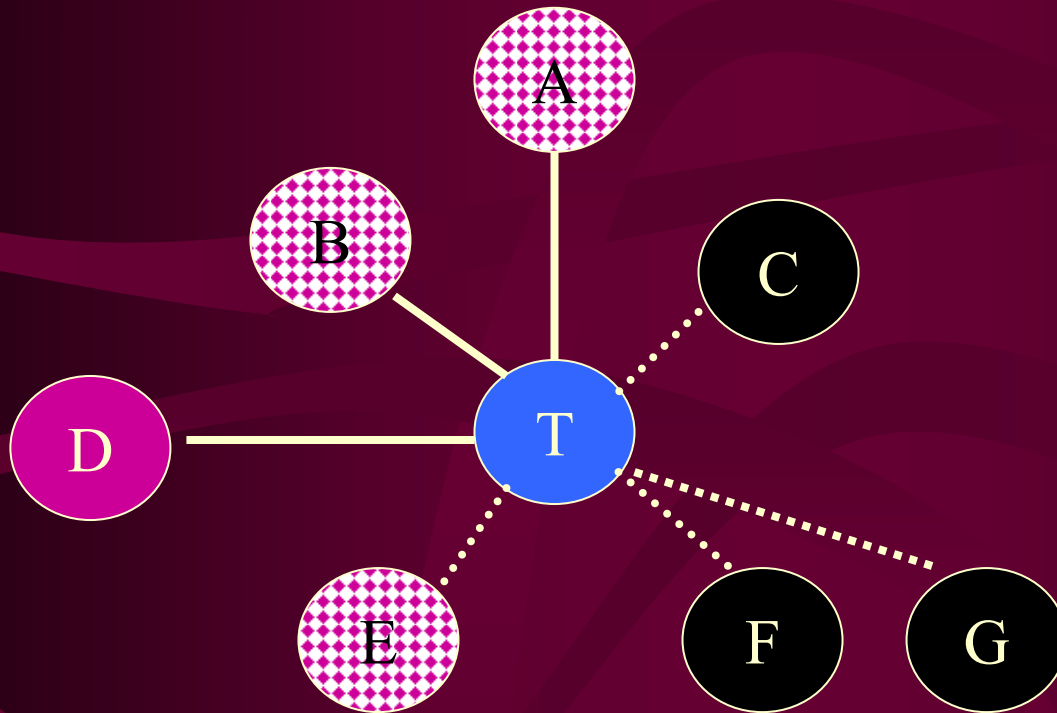
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



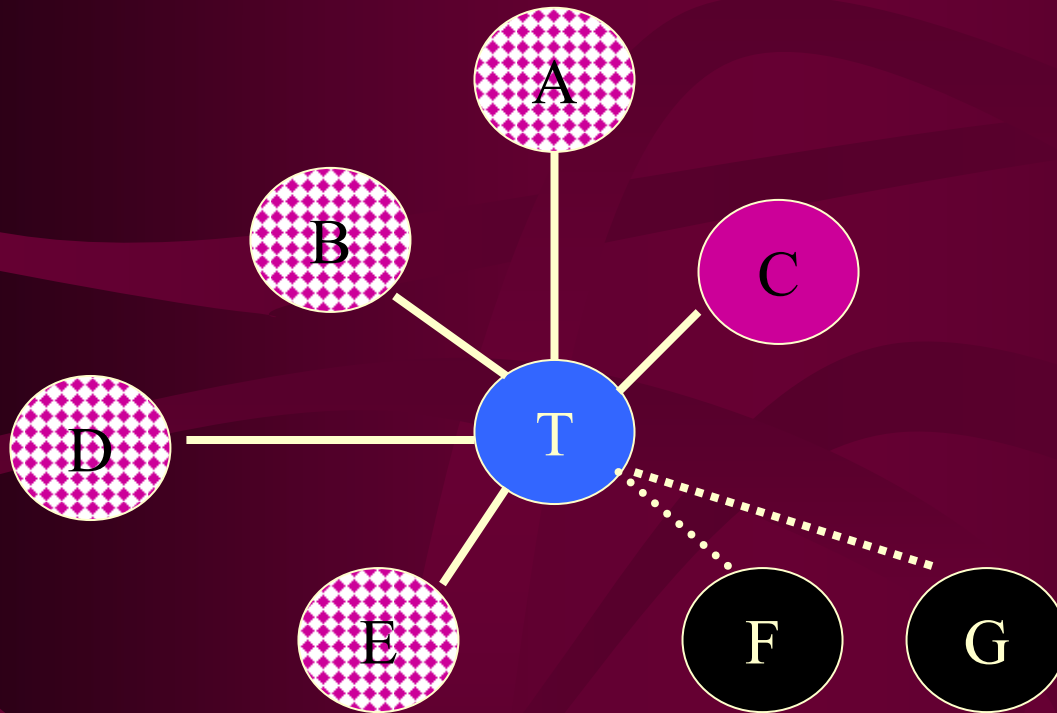
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



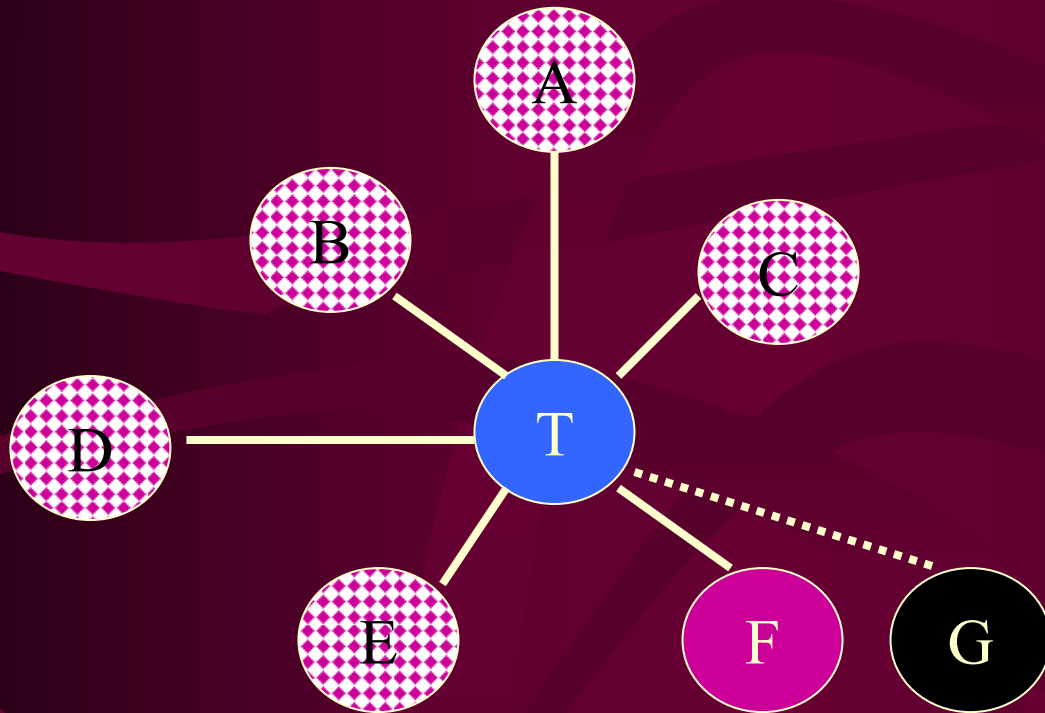
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



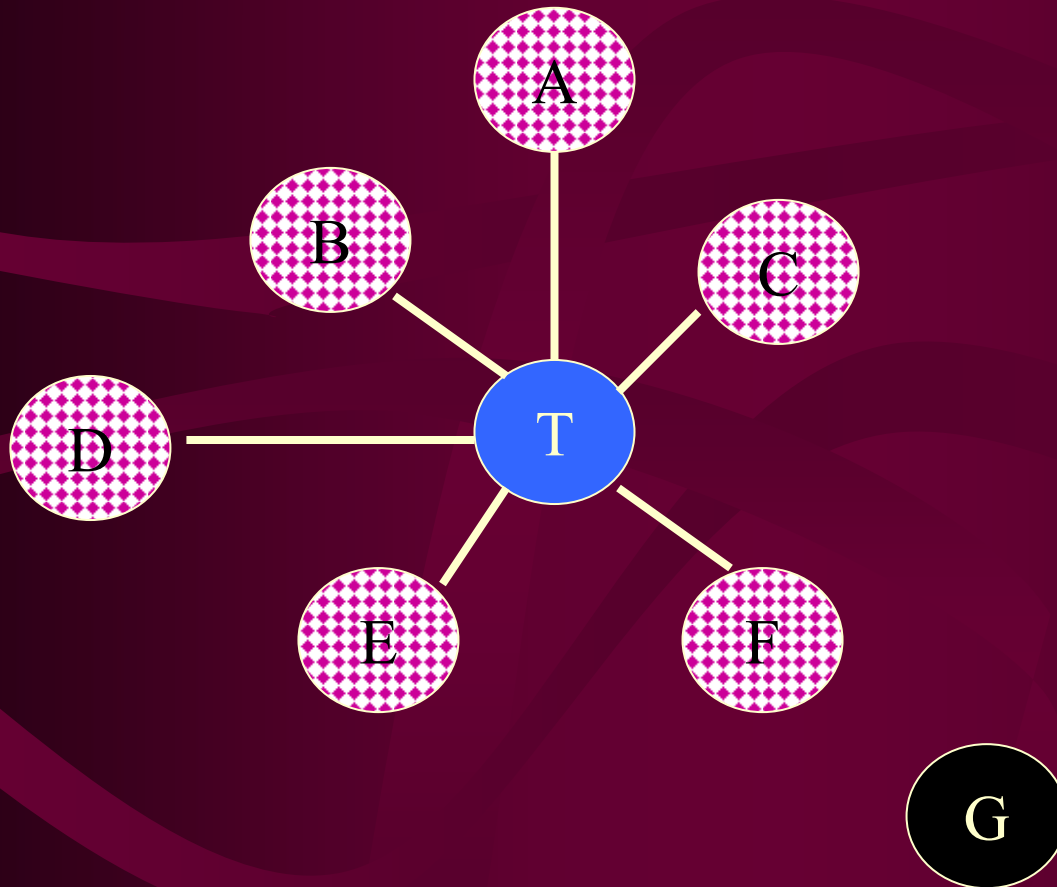
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



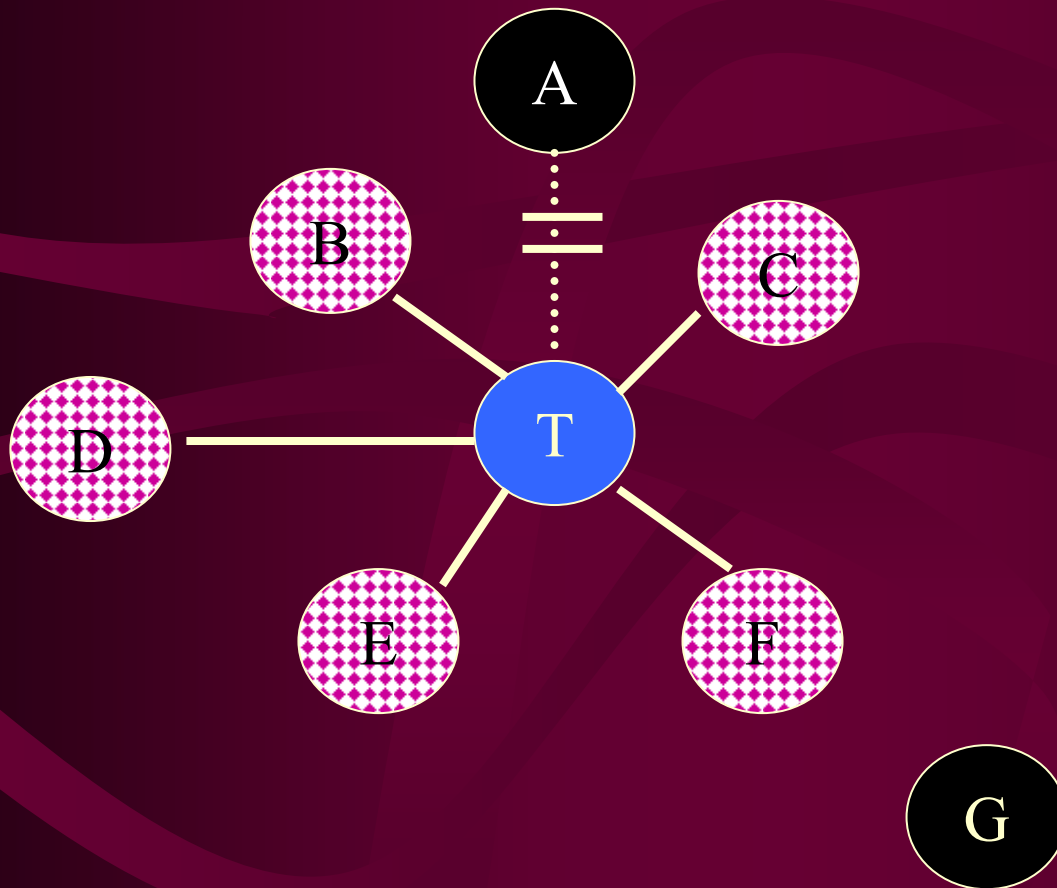
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Small (Relative to the Available Sample)



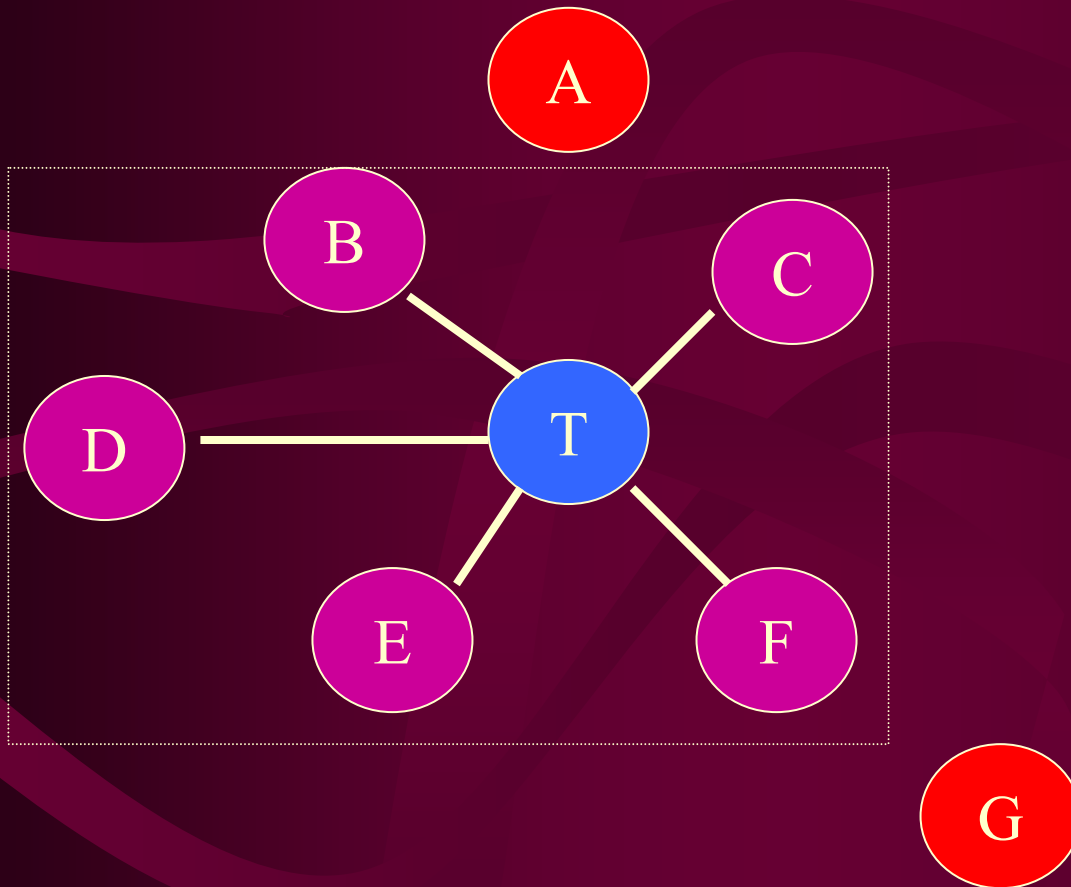
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When $MB(T)$ Is Small (Relative to the Available Sample)



PRUNING PHASE: CONDITIONAL INDEPENDENCE TEST

A Scalable Algorithm for Learning the MB, When $MB(T)$ Is Small (Relative to the Available Sample)



Undirected edges!

A Scalable Algorithm for Learning the MB, When MB(T) Is Large (Relative to the Available Sample)

MMPC/MMMB algorithm (outline)

Start with an empty *CurrentDCE*

Repeat

Find the variable V_i that maximizes $f(V_i, T, \text{CurrentDCE})$,

//where f is a heuristic such that f returns a non-zero value for every variable that is a direct cause or direct effect of T

If $f(V_i, T, \text{CurrentDCE}) > \epsilon$

Admit candidate variable V_i into *CurrentDCE*

Else Exit loop

For all subsets S of *CurrentDCE* of size at most k

Remove any V_i in *CurrentDCE* not in S , such that $I(V_j : T | S)$

Until *false*

If *desired-output=direct-causes-effects*

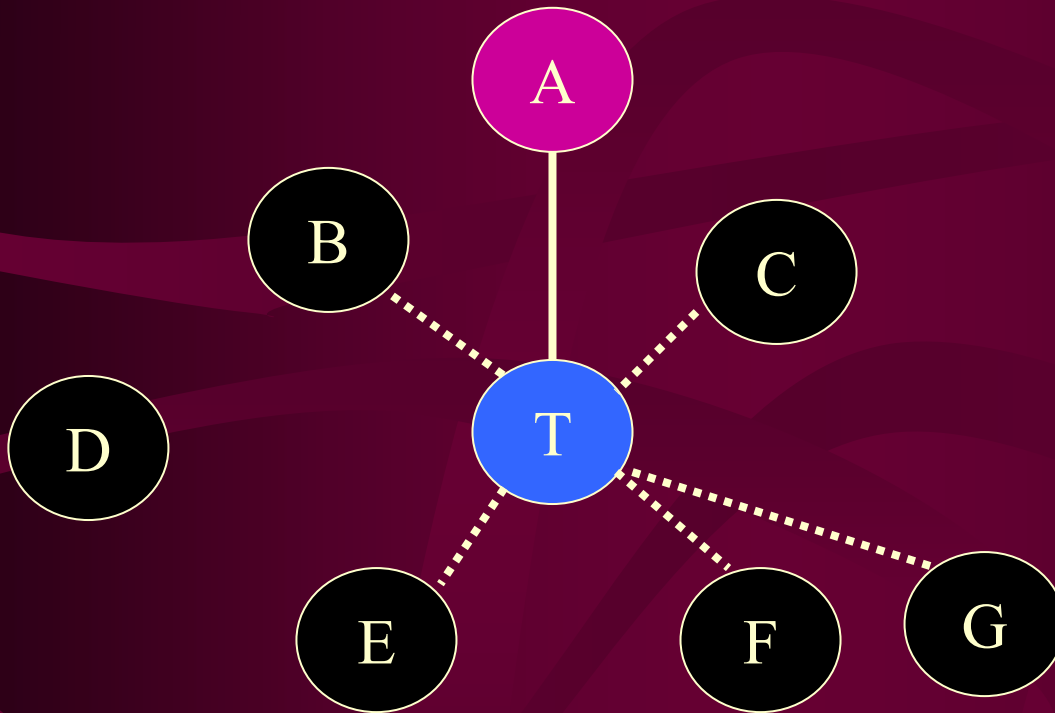
then return *CurrentDCE* //MMPC

else if *desired-output=markov-blanket* //MMMB

then return: $\cup_i (S/\text{LCN}(D, V_i, \text{direct-causes-effects}))$

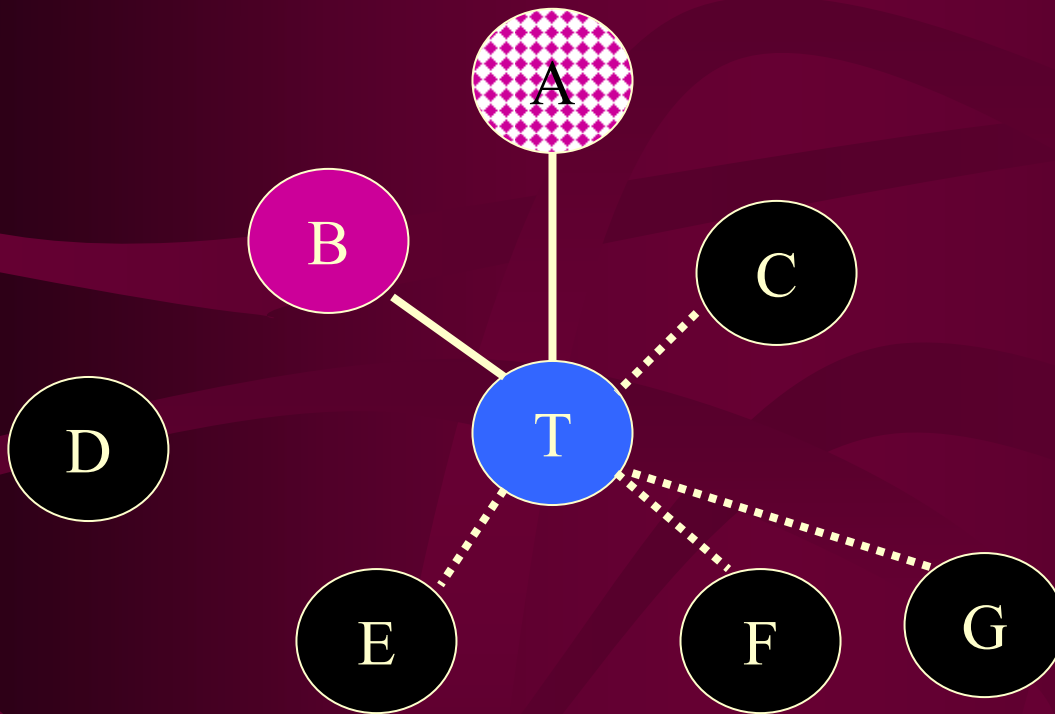
//where V_i is a member of *CurrentDCE*)

A Scalable Algorithm for Learning the MB, When MB(T) Is Large (Relative to the Available Sample)



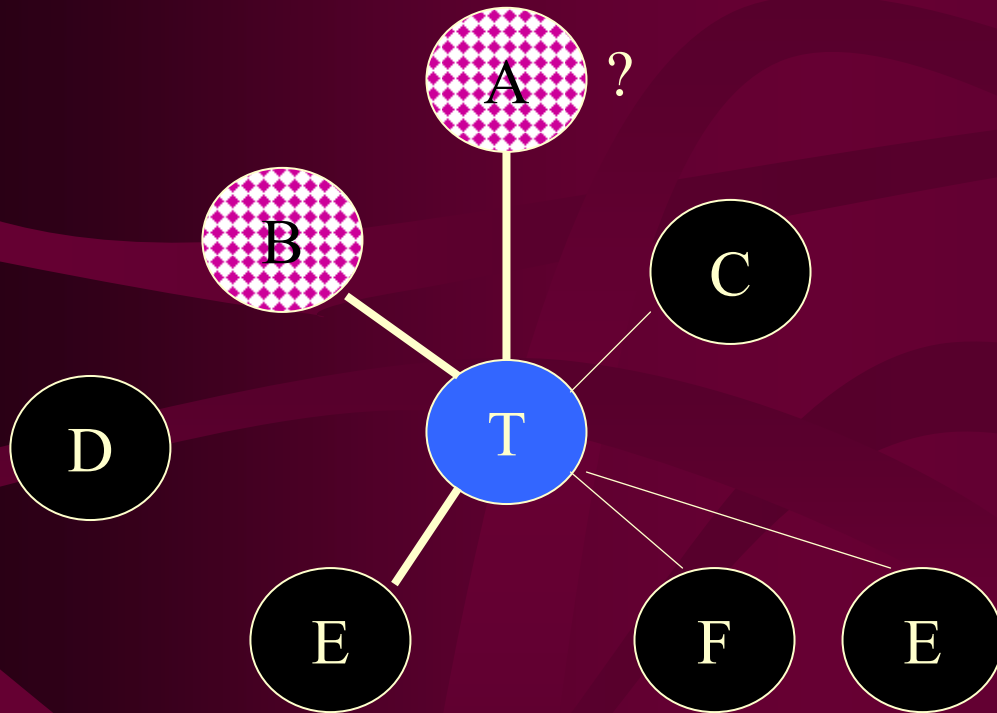
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When MB(T) Is Large (Relative to the Available Sample)



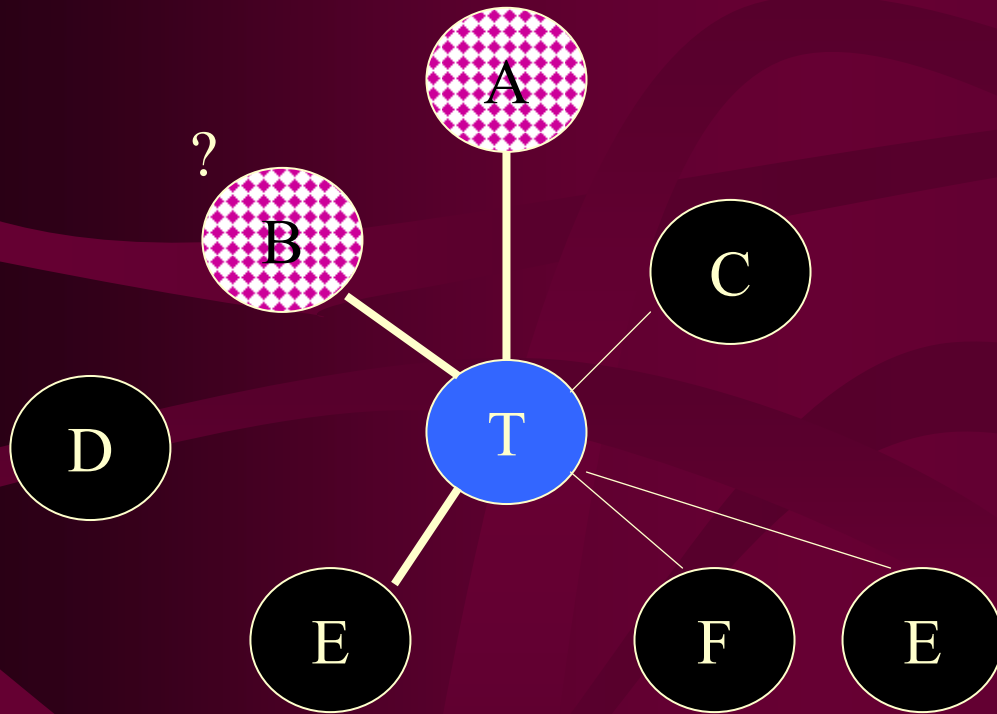
INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

A Scalable Algorithm for Learning the MB, When $MB(T)$ Is Large (Relative to the Available Sample)



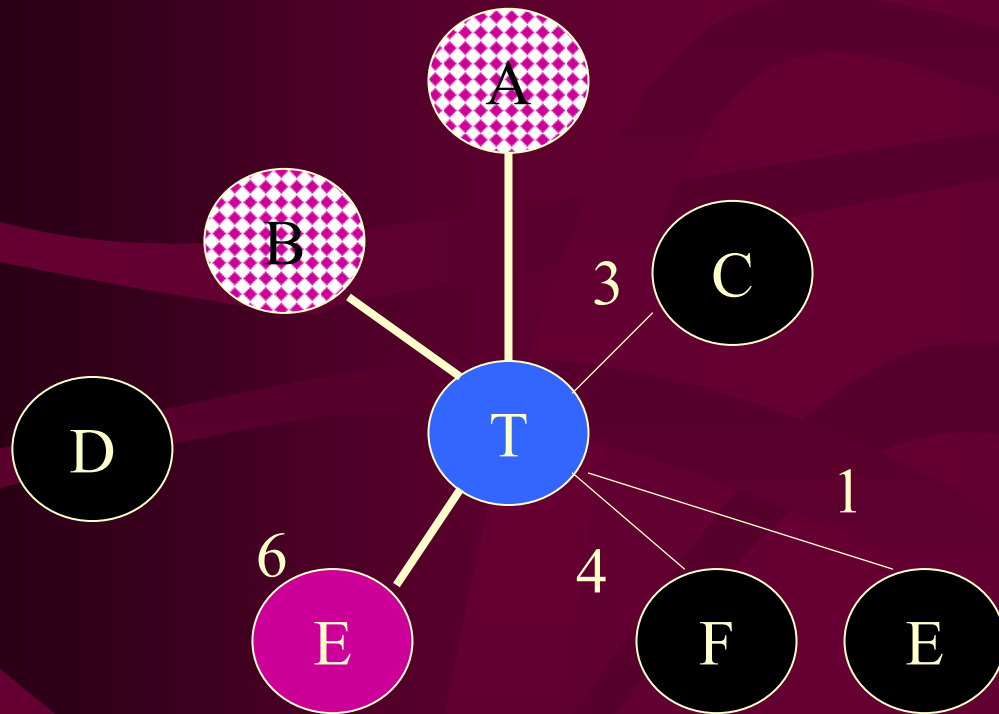
PRUNING: CONDITIONAL INDEPENDENCE TEST

A Scalable Algorithm for Learning the MB, When $MB(T)$ Is Large (Relative to the Available Sample)

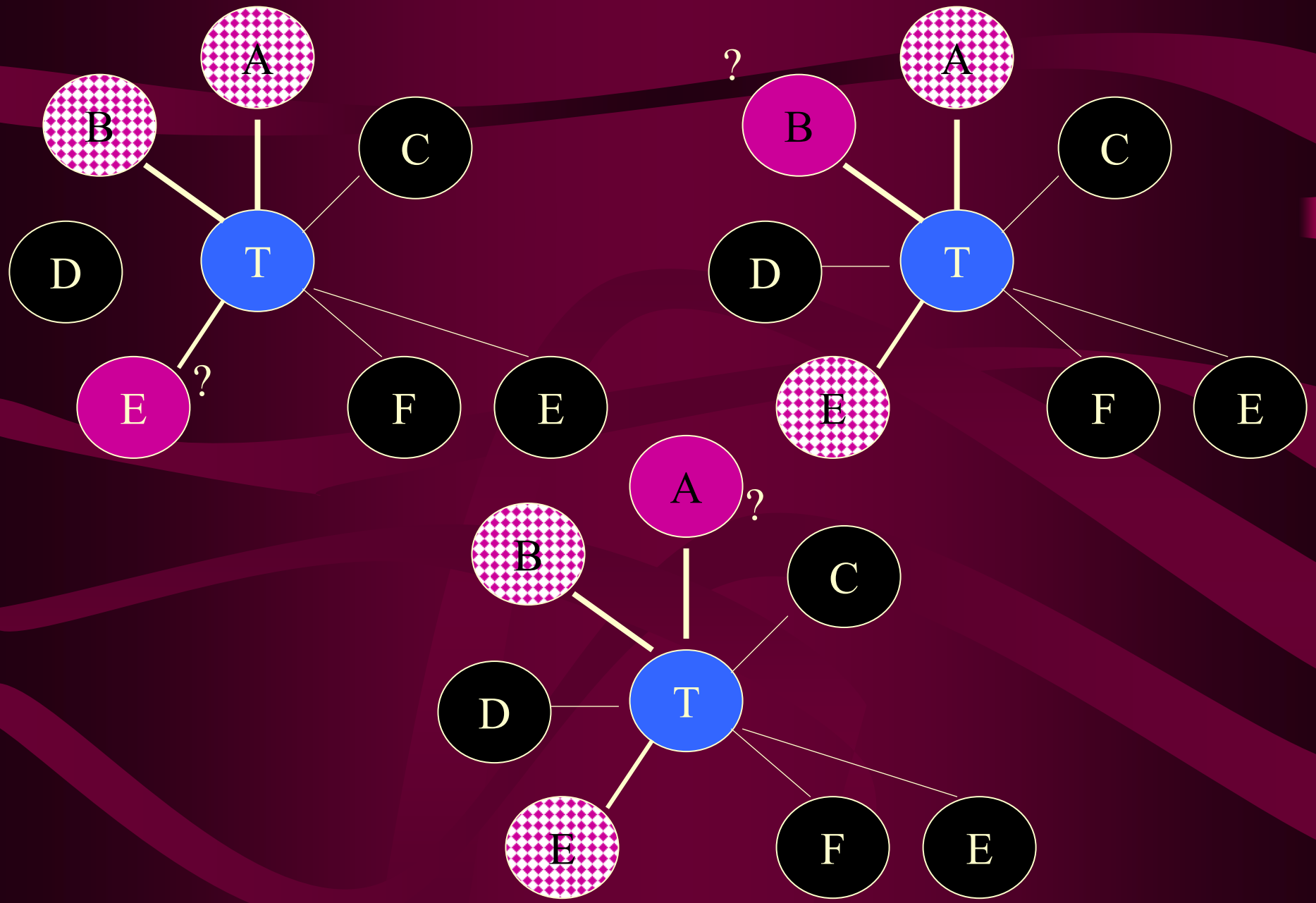


PRUNING: CONDITIONAL INDEPENDENCE TEST

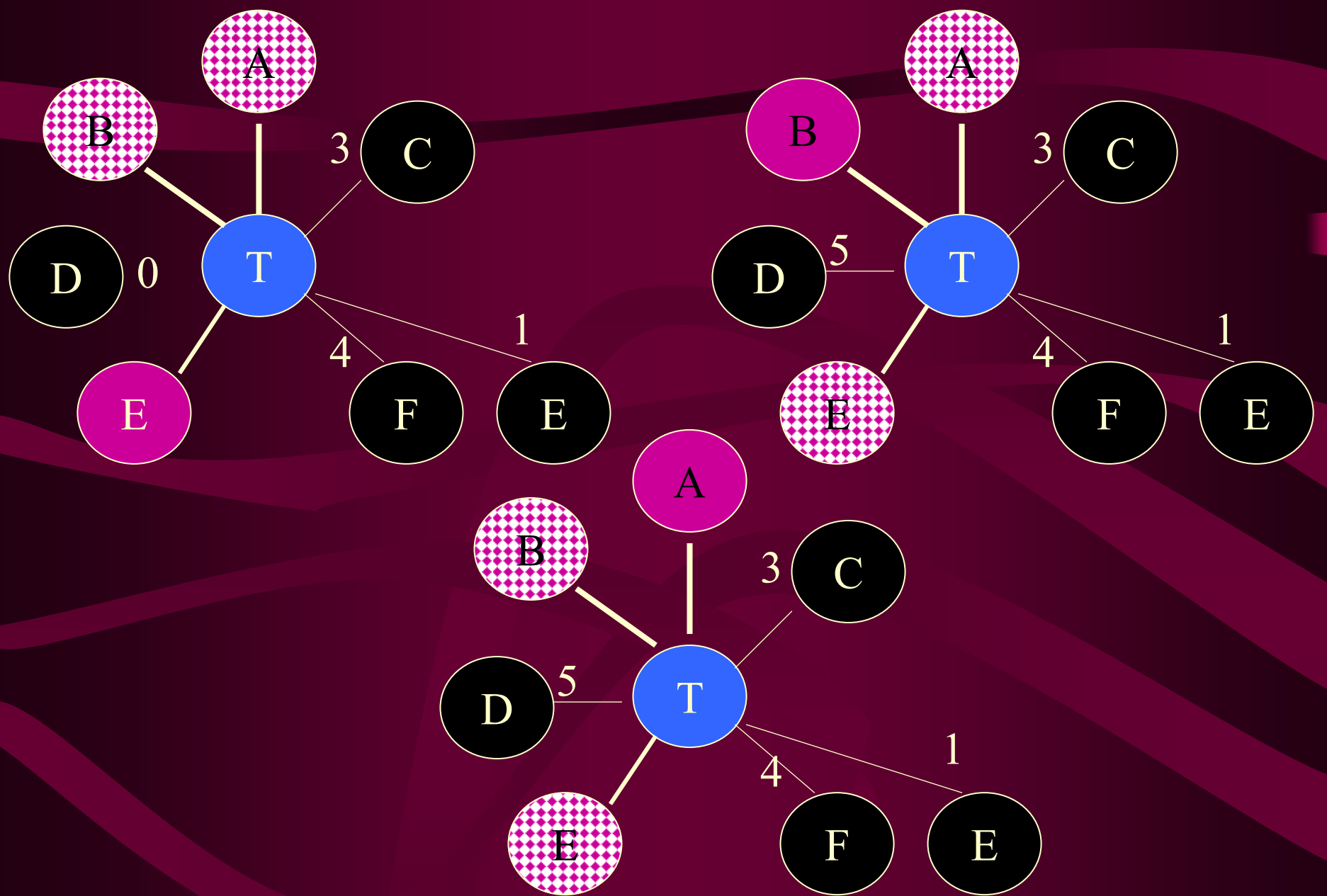
A Scalable Algorithm for Learning the MB, When MB(T) Is Large (Relative to the Available Sample)



INCLUSION PHASE: MAXIMUM CONDITIONAL ASSOCIATION HEURISTIC

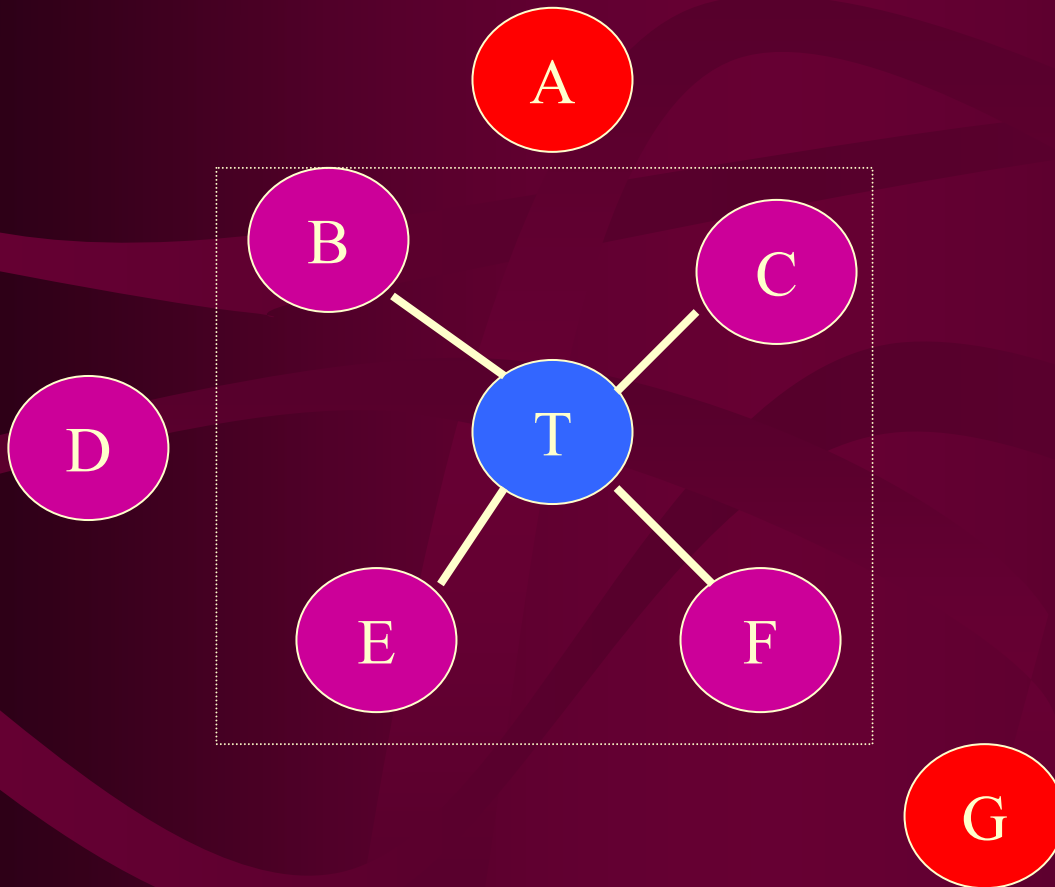


PRUNING: ALL POSSIBLE CONDITIONAL INDEPENDENCE TESTS

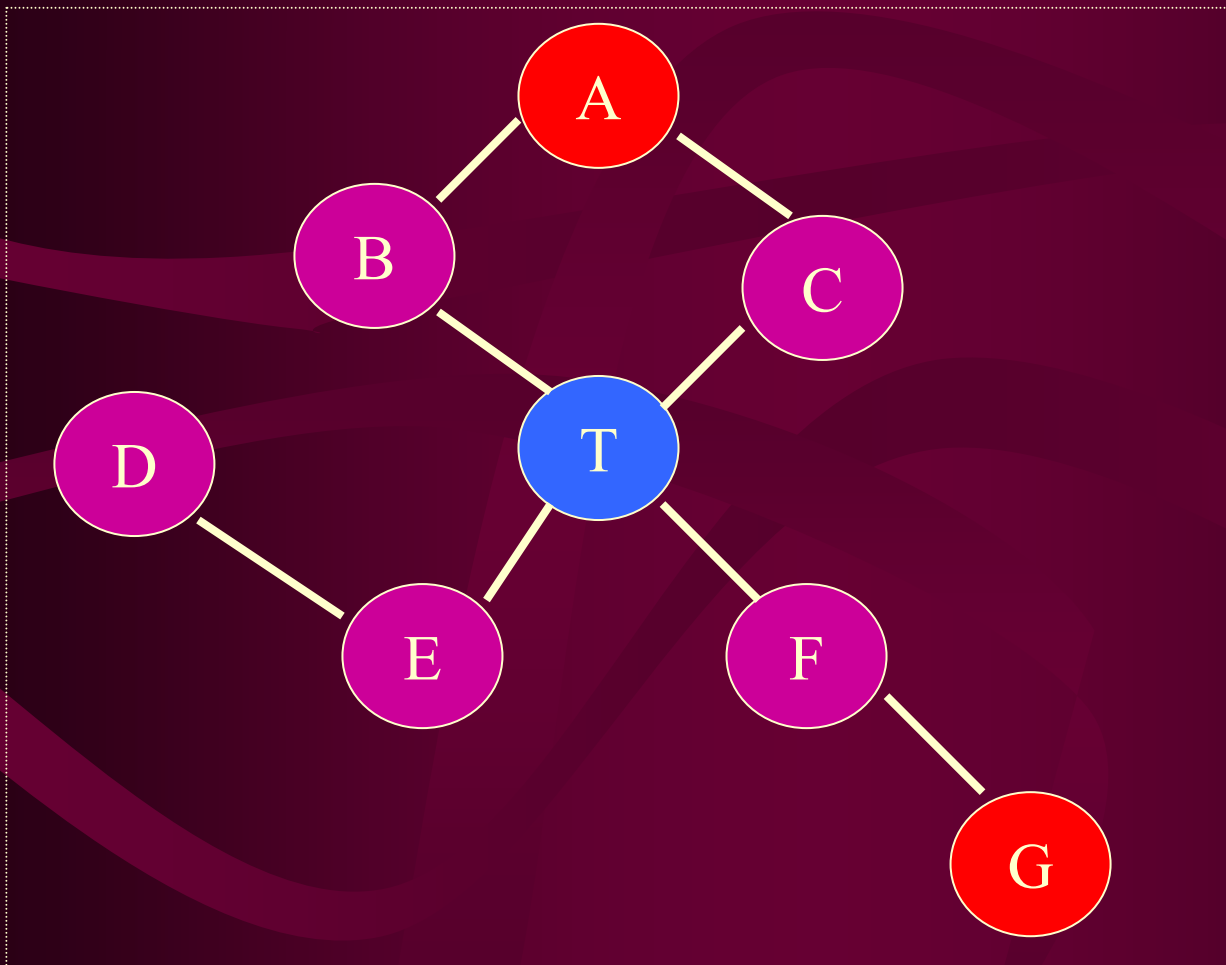


INCLUSION PHASE: MAX-MIN CONDITIONAL ASSOCIATION HEURISTIC

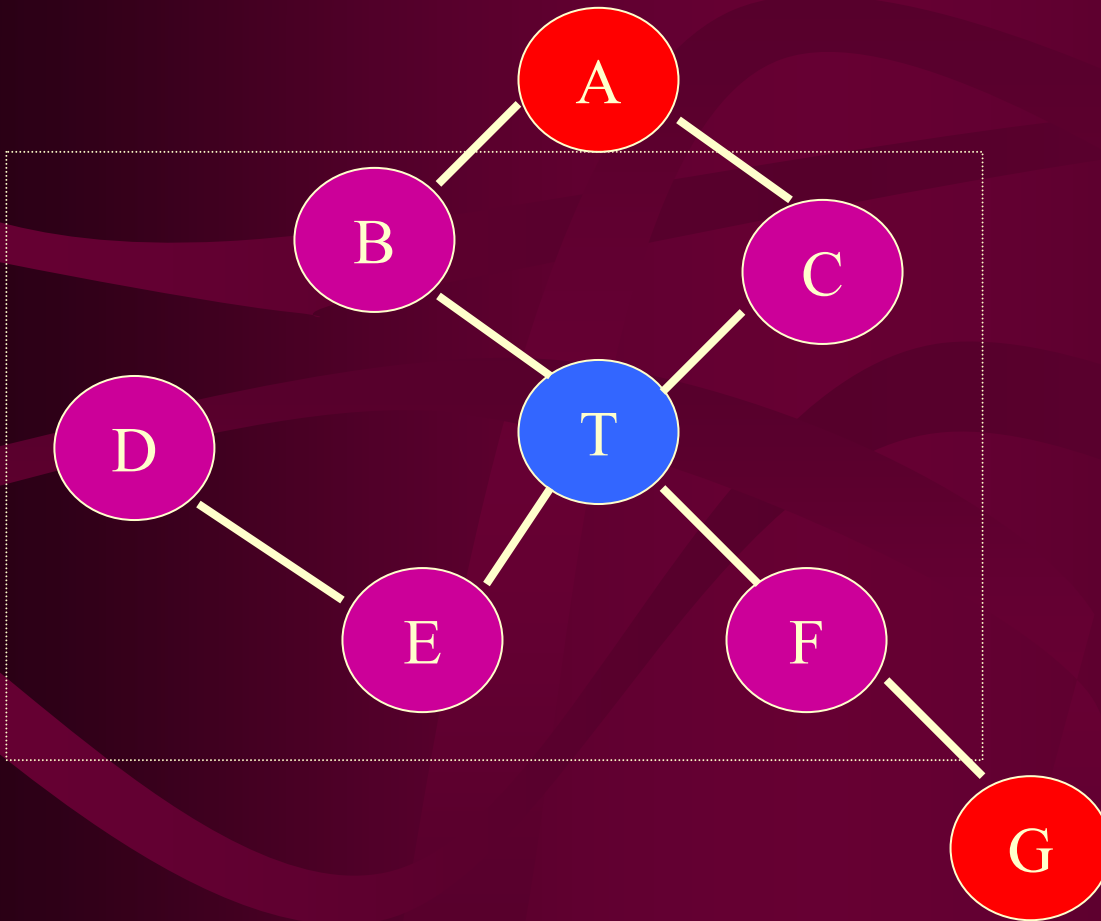
And so on...
until we get all direct causes and direct effects



By recursive application for each direct cause/effect we get a superset of the MB(T)



By Using Additional Pruning Steps we get MB(T) or a tight superset thereof



Biomedical Applications: Thrombin & Lung Cancer

Example application in drug R&D: Task, Data & Methods

- Task: given 139,351 molecular structural properties classify molecules according to whether they bind to thrombin (and thus are good candidates as anti-clotting agents)
- Data from: Cheng et al 2001 [KDD Cup 2001, DuPont Pharmaceuticals]
- Methods: Nested cross-validation: one level to optimize classifier parameters and one level to estimate performance

Thrombin Task: Data Splits

DATA SET	SIZE (% OF split)	ACTIVE	INACTIVE	% OF ACTIVE IN SPLIT
TOTAL	2543 (-)	192	2351	7.6
TRAIN	2000 (78.7)	151	1849	7.6
TRAIN-TRAIN	1300 (65)	90	1210	6.9
VALIDATION	700 (35)	61	639	8.7
TEST	543 (21.3)	41	502	7.6

Thrombin Task: Results

	IAMB (MI, Th=0.0143)	IAMB Chunked (MI, Th=0.0143, 14 chunks)	MMMB (MI, Th=0.01, Cond =5)	UAF	RFE	All
LSVM	88.3000%	93.4800%	94.8000%	94.7300%	93.2878%	93.4300%
PSVM	86.2600%	93.7800%	93.4400%	94.4600%	92.4716%	93.6900%
SBC	92.7500%	93.2500%	94.7700%	94.0500%	85.2128%	80.3300%
KNN	94.0600%	91.2100%	93.7900%	94.7800%	89.7095%	88.2100%
NN	94.1500%	93.3000%	93.5900%	88.8900%	92.0416%	N/A
Averages	91.1040%	93.0040%	94.0780%	93.3820%	90.5447%	88.9150%
Number of features	8	9	27	200	8709	139351

Thrombin Task: Parallelization

ALGORITHM	TIME (HRS)	NOTES
IAMB	6.7	1 CPU
PARALLEL IAMB	0.5	14 CPUs
PARALLEL CHUNKED IAMB	0.5	14 CPUs

Time results from parallel vs regular versions of our algorithms on the Thrombin data set (Platform: 128 M RAM, 600MHz PIII).

Thrombin Task: Time Efficiency

- MMPC: 3.1 hours
- MMMB: 15.3 hours
- IAMB: 2.9 hours
- IAMB Chunked: 2.6 hours
- Platform: Unoptimized, interpreted Matlab code on a Pentium4, 2 GHz, Windows 2000

Lung Cancer: Tasks

- Diagnose Normal vs Cancer cells
- Diagnose AdenoCa vs Squamous Ca
- Diagnose Metastatic vs Non-Metastatic Lung Cancers
- Predict the values of individual genes & recover their local causal neighborhoods (i.e., direct causes and direct effects)

Lung Cancer: Data & Methods

- Bhattacharjee et al. PNAS, 2001
- 12,600 gene expression measurements obtained using Affymetrix oligonucleotide arrays
- 203 patients and normal subjects, 5 disease types, (plus staging and survival information)
- Nested cross-validation: one level to optimize classifier parameters and one level to estimate performance

Lung Cancer: Distinguishing Normal vs Cancer Cells [Model Performance]

17 non-cancerous patients (186 cancerous), 5-fold cross-validation (3(5) non-cancerous patients per fold)

	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	98.62%	99.28%	99.73%	97.03%	99.26%	99.64%
PSVM	99.55%	99.19%	99.46%	97.48%	99.26%	99.64%
KNN	93.05%	93.19%	92.74%	87.83%	97.33%	98.11%
NN	100.00%	100.00%	99.91%	97.57%	99.80%	N/A
Averages of FS Algorithms	97.80%	97.91%	97.96%	94.97%	98.91%	99.13%

Lung Cancer: Distinguishing Normal vs Cancer Cells [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	1	19	18
MMMB	84		9	103	101
MMMB+PC	76	0		94	92
RFE	6	6	6		2
UAF	99	98	98	96	

Lung Cancer: Distinguishing Normal vs Cancer Cells [Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	19	0
	MMMB	103	1
	MMMB+PC	94	1
	RFE	6	0
	UAF	100	1

- p53 and p63 a strong homolog to p53, - p16 (cyclin-dependent kinase inhibitor 2A that inhibits CDK4))
- k-ras
- akt (v-akt murine thymoma viral oncogene homolog 1; AKT1 , AKT2 ,)
- hTERT (telomere reverse transcriptase)
- c-myc (v-myc avian myelocytomatosis viral oncogene homolog)
- ornithine decarboxylase 1, - kallikrein 11,
- surfactant protein (surfactant protein A binding protein, surfactant, pulmonary-associated protein D, surfactant, pulmonary-associated protein C, surfactant, pulmonary-associated protein B)

Lung Cancer: Distinguishing AdenoCa vs Squamous Cancer [Model Performance]

21 squamous patients (139 adeno), 5-fold cross-validation (4(5) squamous patients per fold)

	MMPC	MME	MME+PC	RFE	UAF Cross-Validated	All Features
LSVM	99.09%	98.49%	98.49%	98.57%	99.32%	98.98%
PSVM	96.91%	98.72%	98.72%	98.57%	98.70%	99.07%
KNN	98.26%	93.07%	93.07%	91.49%	95.57%	97.59%
NN	98.88%	99.32%	99.38%	98.70%	99.63%	N/A
Averages of FS Algorithms	98.28%	97.40%	97.41%	96.83%	98.30%	98.55%

Lung Cancer: Distinguishing AdenoCa vs Squamous Cancer [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	0	12	2
MMMB	52		0	64	37
MMMB+PC	52	0		64	37
RFE	11	11	11		5
UAF	489	472	472	493	

Lung Cancer: Distinguishing AdenoCa vs Squamus Cancer

[Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	13	0
	MMMB	65	0
	MMMB+PC	65	0
	RFE	12	0
	UAF	500	2

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Model Performance]

7 metastatic patients (132 non-metastatic), 7-fold cross-validation (1 metastatic patient per fold)

	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	90.87%	97.32%	93.35%	96.43%	95.63%	96.83%
PSVM	88.89%	97.62%	94.84%	97.62%	96.43%	96.33%
KNN	86.90%	84.92%	84.92%	92.46%	89.29%	92.56%
NN	90.08%	96.03%	96.83%	96.83%	86.90%	N/A
Averages of FS Algorithms	89.18%	93.97%	92.49%	95.84%	92.06%	95.24%

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Relative Novelty of Genes]

Contributed by (vertical) compared with (horizontal)	MMPC	MMMB	MMMB+PC	RFE	UAF
MMPC		0	0	14	6
MMMB	62		5	75	55
MMMB+PC	57	0		71	53
RFE	6	5	6		2
UAF	492	479	482	496	

Lung Cancer: Distinguishing Metastatic vs Non-Metastatic Cancers [Size of feature Sets & Literature Novelty of Genes]

Final Model	Feature Selection Method	Number of features discovered	Number of intersections with gene list
	MMPC	14	0
	MMMB	76	0
	MMMB+PC	71	0
	RFE	6	0
	UAF	500	3

Lung Cancer Models: Time Efficiency

- MMPC: 0.5 to 5 minutes (depending on the data split of the cross-validation)
- MMMB: 15 to 70 minutes (depending on the data split)
- Platform: Unoptimized, interpreted Matlab code on a Pentium 4, 2 GHz, Windows 2000

Biological Meaning?

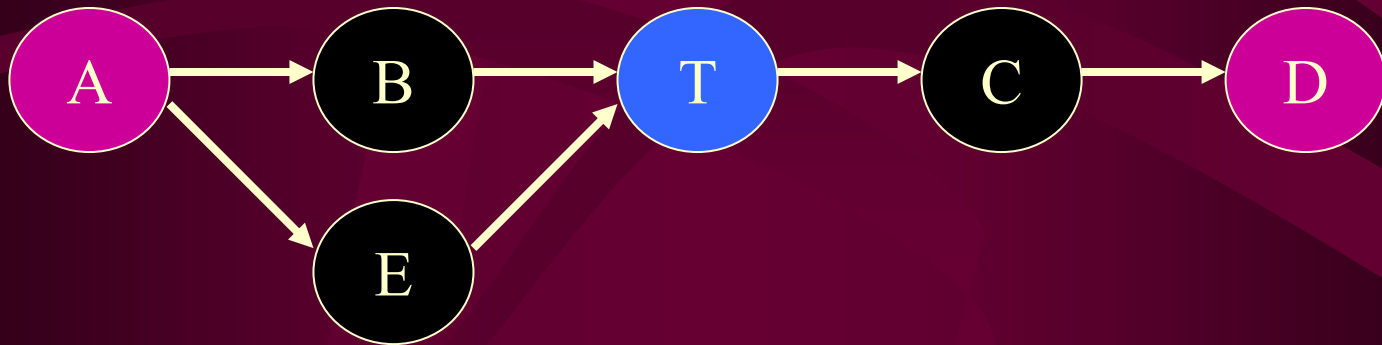
- Both our methods and the baseline ones produce high-quality diagnostic models that may prove to be of significant clinical value.
- The selected features are radically different however!
- Selected genes are also novel relative to genes of known importance for all methods!

Biological Meaning?

- Problem: all methods produce good predictor sets of biological states. *Why are these sets so different?*
- Interpretation:
 - Different methods have different *inductive biases* (i.e., preferences for one type of model over the rest) and are optimized for different tasks.
 - We need to interpret algorithmic results in the context of the inductive biases of each method.

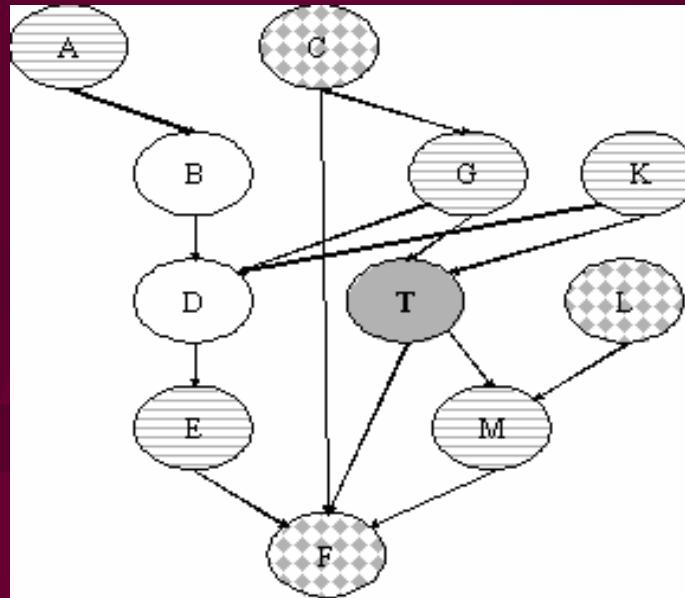
Characterizing the Genes Selected By Various Methods

- Claim: By Design Markov Blanket/Local causal neighborhood methods output genes that are causally “closer” to the target variable
- A Method to test the claim: **Relative Conditional Blocking**
(intuitive example: $\{A,D\}$ may predict T as well as the larger $\{B,C,E\}$ set, however $\{B,C,E\}$ are closer to T than $\{A,D\}$)



Note: An alternative more technical way to describe this measure is “Relative Divergence from the Causal Markov Condition”; It answers the “causal proximity” question

Characterizing the Genes Selected By Various Methods



Suppose algorithm *Alg1* returns variables $V_1 = \{A, G, K, E, M\}$ (textured with horizontal lines) and algorithm *Alg2*, variables $V_2 = \{C, F, L\}$ (with checkered texture). Conditioned on some subset of V_1 variables L and C are independent of the target (66% of V_2 is “blocked” by V_1). Conditioned on some subset of V_2 variables A and E are independent of T (40% of V_1 is “blocked” by V_2). We conclude that *Alg1*’s output is closer to the direct causes and effects of T , than *Alg2*’s output.

Relative Conditional Blocking In Lung Cancer data

<i>Percentage of features 1 (vertical) eliminated by features 2 (horizontal)</i>	MMPC	RFE
Task 1: Non-Metastatic Adeno vs. Metastatic Adeno		
MMPC	-1	42.86%
RFE	83.33%	-1
Task 2: Cancerous vs. Non-Cancerous		
MMPC	-1	10.53%
RFE	100.00%	-1
Task 3: Adeno vs. Squamous		
MMPC	-1	16.67%
RFE	100.00%	-1

Predicting Single Genes: Performance

(Tumor protein 63 kDa with strong homology to p53)

69 patients with over-expressed gene (134 under-expressed cases), 6-fold cross-validation (11(14) over/expr cases/fold)

	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	94.21%	91.96%	92.74%	80.42%	87.92%	90.32%
PSVM	89.67%	89.98%	91.64%	75.35%	87.93%	84.93%
KNN	89.48%	88.68%	92.43%	79.27%	88.29%	84.98%
NN	93.38%	93.78%	93.90%	80.39%	89.27%	N/A
Averages of FS A	91.69%	91.10%	92.68%	78.86%	88.35%	86.74%

Predicting Single Genes: Performance

(V-AKT MURINE THYMOMA VIRAL ONCOGENE HOMOLOG 1; AKT1)

69 patients with over-expressed gene (134 under-expressed cases), 6-fold cross-validation (11(14) over-expressed cases per fold)

	MMPC	MMMB	MMMB+PC	RFE	UAF Cross-Validated	All Features
LSVM	68.15%	70.30%	79.71%	73.83%	73.22%	70.00%
PSVM	68.43%	72.81%	77.71%	71.53%	66.29%	68.39%
KNN	68.99%	72.74%	73.78%	72.58%	71.11%	76.29%
NN	71.86%	78.29%	80.70%	71.78%	71.52%	N/A
Averages of FS Algorithms	69.36%	73.53%	77.98%	72.43%	70.54%	71.56%

Predicting Single Genes: Stability, Novelty, Relative Blocking

- Results same when choosing different thresholds and when doing regression instead of classification
- Results similar in terms of non-intersection of gene lists among methods and with gene list
- Relative blocking same as with prediction of disease
- Size of direct causes/effects: 7 and 9 (compared to 14, 19, and 13 for three diagnostic tasks)

Conclusions

1. By using novel Markov Blanket Methods we can derive optimal or near-optimal prediction and diagnostic models in gene expression and structural biology tasks
2. The Markov Blanket methods achieve excellent to outstanding reduction in the number of predictors
3. The Markov Blanket methods are efficient in time and sample even in the presence of massive numbers of variables
4. The Markov Blanket methods offer additional information than state-of-the art methods and have a causally-oriented interpretation

A Sampler of Questions & Future Research Directions

- **Methods:**
 - Improved pruning for MMMB
 - Quality of models when sample is reduced
- **Lung Cancer:**
 - Analysis of all important genes
 - Experimental verification in cell lines
 - Survival models
 - Stability of model performance across datasets
 - Stability of gene sets across datasets
 - Comparison to aCGH
 - Markov Blanket-based molecular subtyping/clustering
- **Other biomedical domains:**
 - Applying these methods to proteomics and treatment response prediction