On Pruning with the MDL Score

Eunice Yuh-Jie Chen, Arthur Choi and Adnan Darwiche

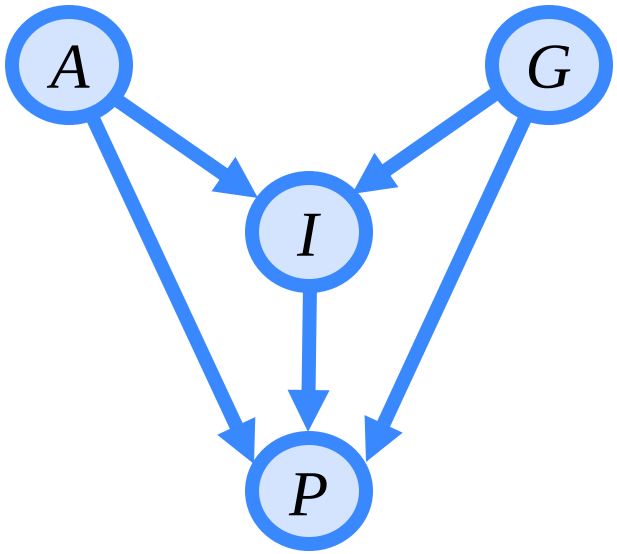
Computer Science Department
UCLA

Bayesian Network Structure Learning

INPUT: Complete Dataset

income	age	gender	payment
40,000	22	M	true
68,000	35	F	false
53,000	25	F	true
85,000	30	F	false
62,000	40	M	false

OUTPUT: Optimal Structure

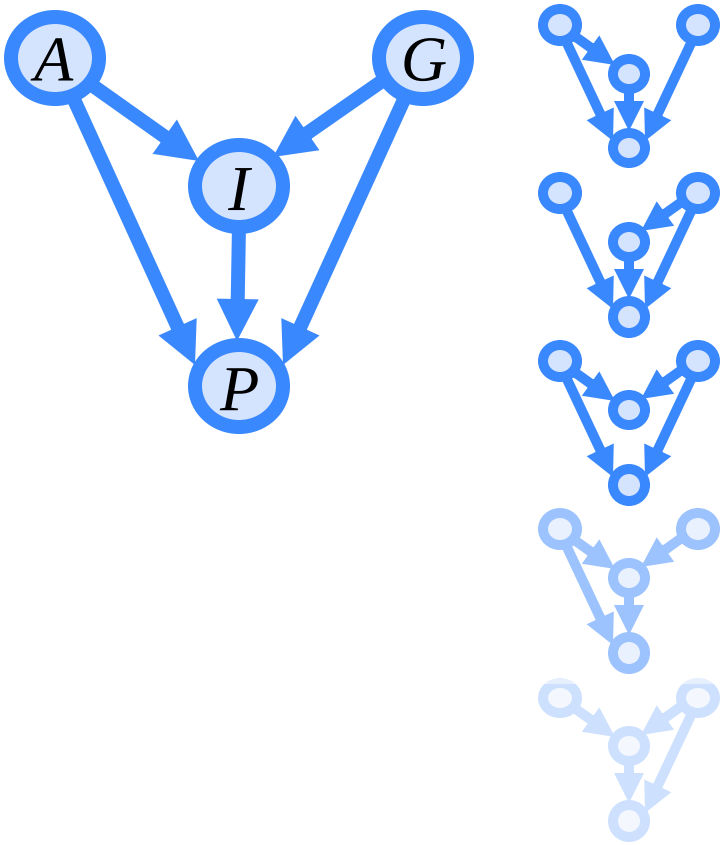


Bayesian Network Enumeration

INPUT: Complete Dataset

income	age	gender	payment
40,000	22	M	true
68,000	35	F	false
53,000	25	F	true
85,000	30	F	false
62,000	40	M	false

OUTPUT: *k*-Best Structures

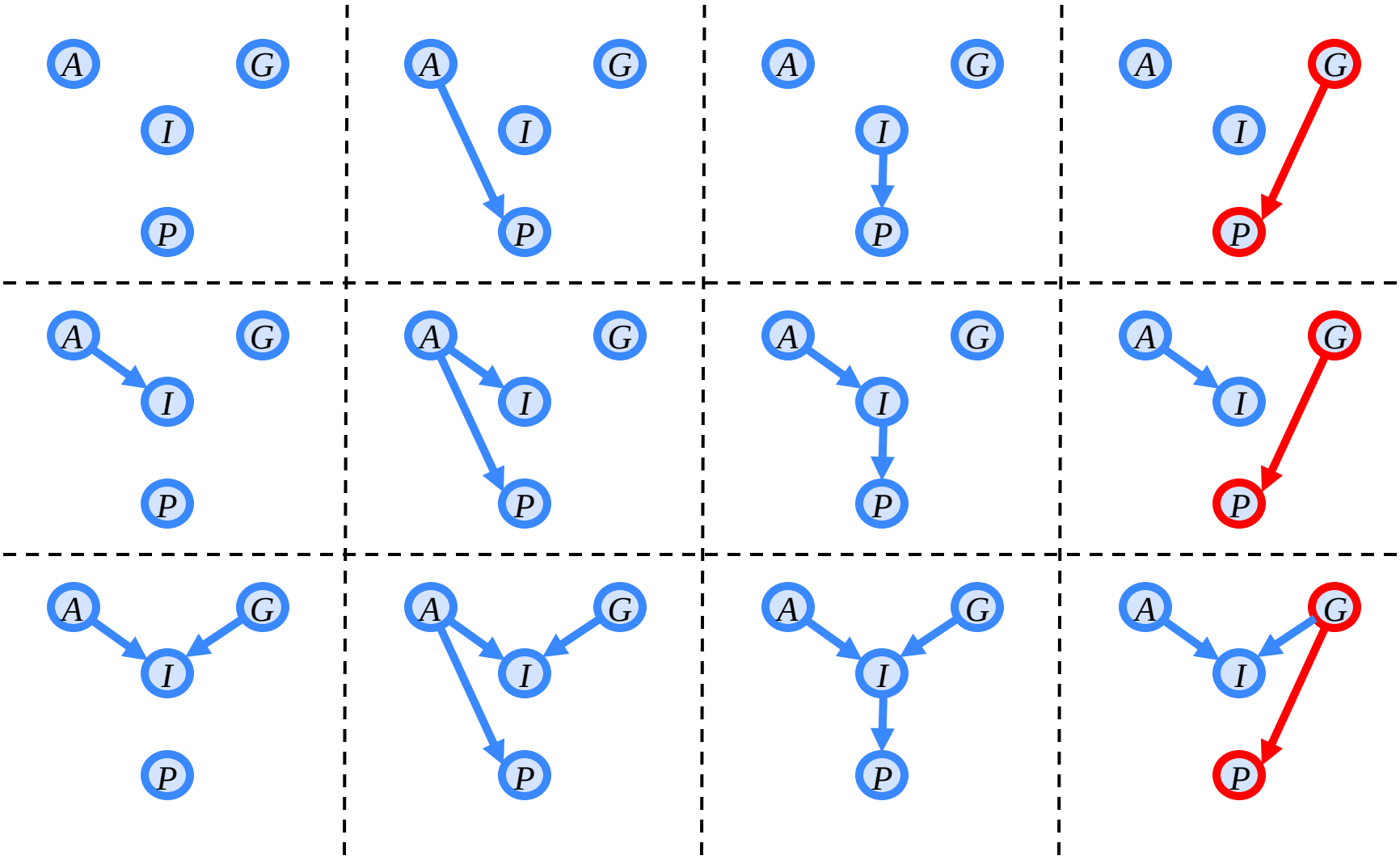


Enumerating BNs: This Talk

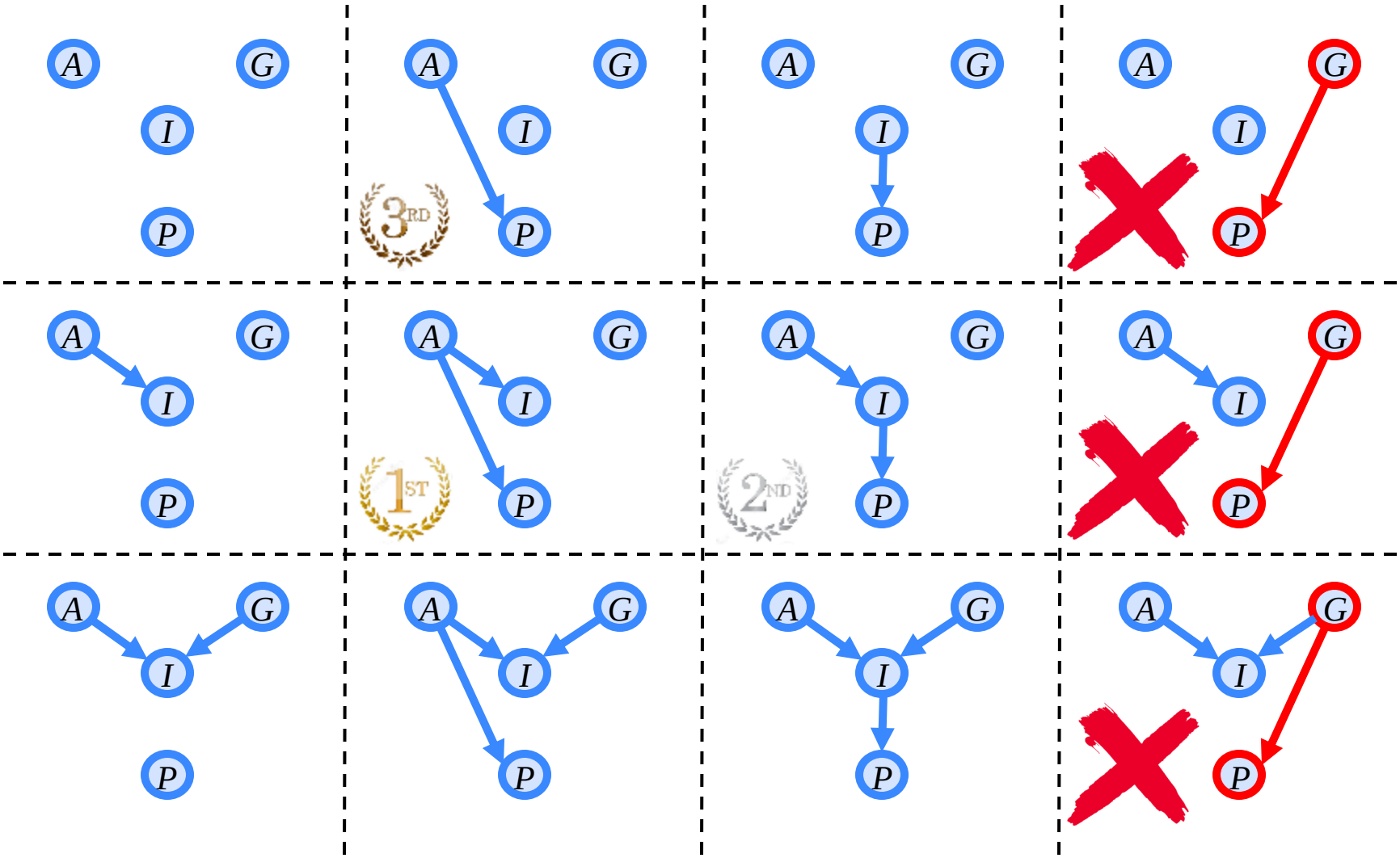
paper	number of variables	search space size
Tian, He & Ram 2010	17	6.27×10^{52}
Chen, Choi & Darwiche 2015	23	6.97×10^{94}
THIS PAPER	**29**	**2.51×10^{148}**

see also [Cussens, Bartlett, Jones & Sheehan 2013], for pedigrees

Scaling Structure Learning: Pruning



Scaling Structure Learning: Pruning



Scaling Structure Learning: Pruning

- **Heuristic Search** (e.g., A*):
  prune search space over DAGs
- **Dynamic Programming**:
  prune sub-problems
- **Integer Linear Prog**:
  reduce the # of ILP variables
- **Score Caching**:
  pruning saves time and memory

Scoring Functions: MDL

$$\text{MDL}(G \mid \mathcal{D}) = -\log Pr(\mathcal{D} \mid \theta) + \frac{K(G)}{2} \log N$$

Score how well the model fits the data

Penalize the complexity of the model

where $K(G)$ is the # of free parameters in DAG G

Scoring Functions: MDL

$$\mathrm{MDL}(G \mid \mathcal{D}) = \sum_{X\mathbf{U}} \mathrm{MDL}(X \mid \mathbf{U})$$

$$= \sum_{X\mathbf{U}} H(X|\mathbf{U}) + \frac{K(X \mid \mathbf{U})}{2} \log N$$

Score how well the family fits the data

Penalize the complexity of the CPT

where $K(X|\mathbf{U})$ is the # of free parameters in the CPT
and $H(X|\mathbf{U})$ is entropy of the empirical distribution

Scoring Functions: MDL

$$\mathsf{MDL}(G \mid \mathcal{D}) = \sum_{X\mathbf{U}} \mathsf{MDL}(X \mid \mathbf{U})$$

$$= \sum_{X\mathbf{U}} H(X|\mathbf{U}) + \frac{K(X \mid \mathbf{U})}{2} \log N$$

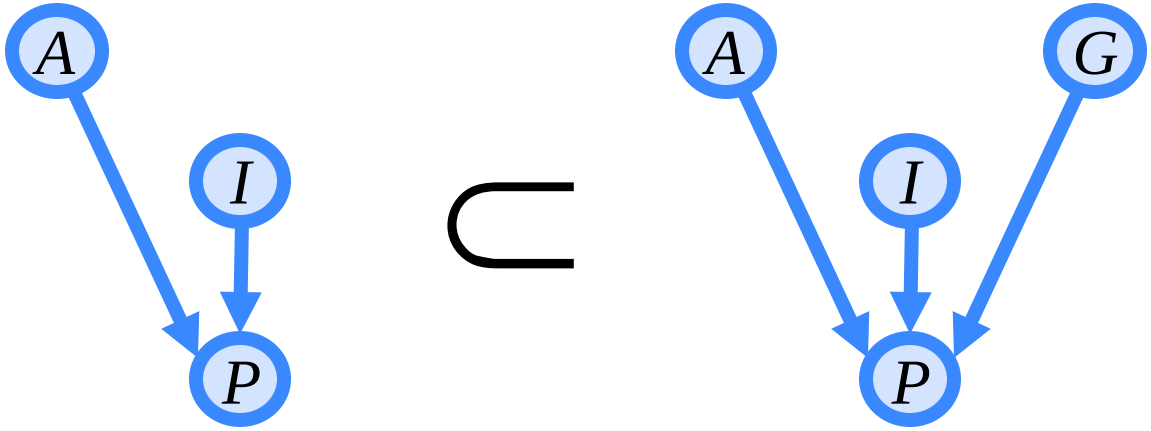
Penalty of large family does not make up for better fit of data:
prune large families

Pruning Rules for MDL

- **[Suzuki 1996, Tian 2000, De Campos and Ji 2011]** Under the MDL score, do not consider families *X***U** where:

$$|\mathbf{U}| > \left\lfloor \log_2 \frac{2N}{\log_2 N} \right\rfloor$$

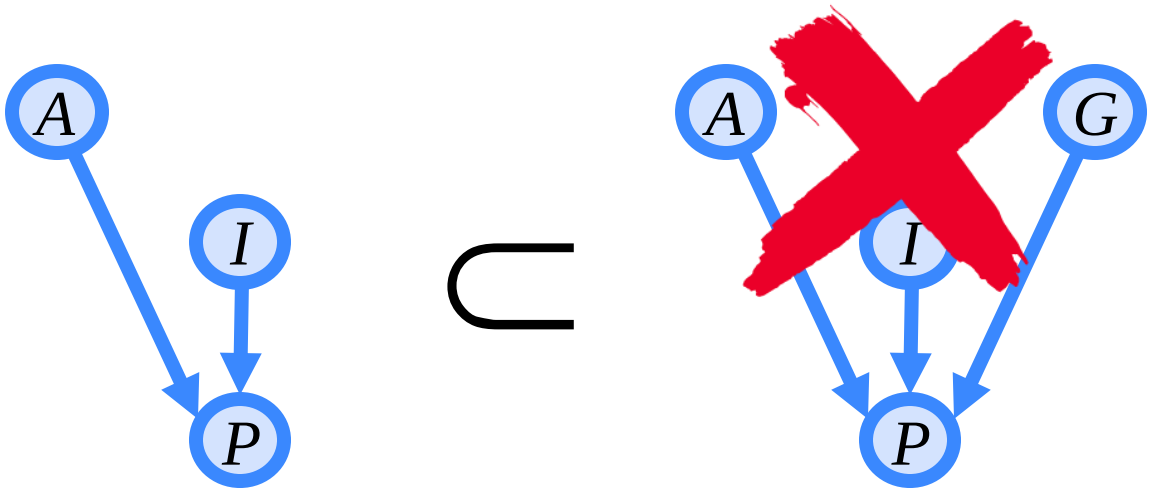
Pruning Rule: Basic Idea



SCORE(P | AI) $<$ **SCORE(P | AIG)**

[Teyssier & Koller 2005]

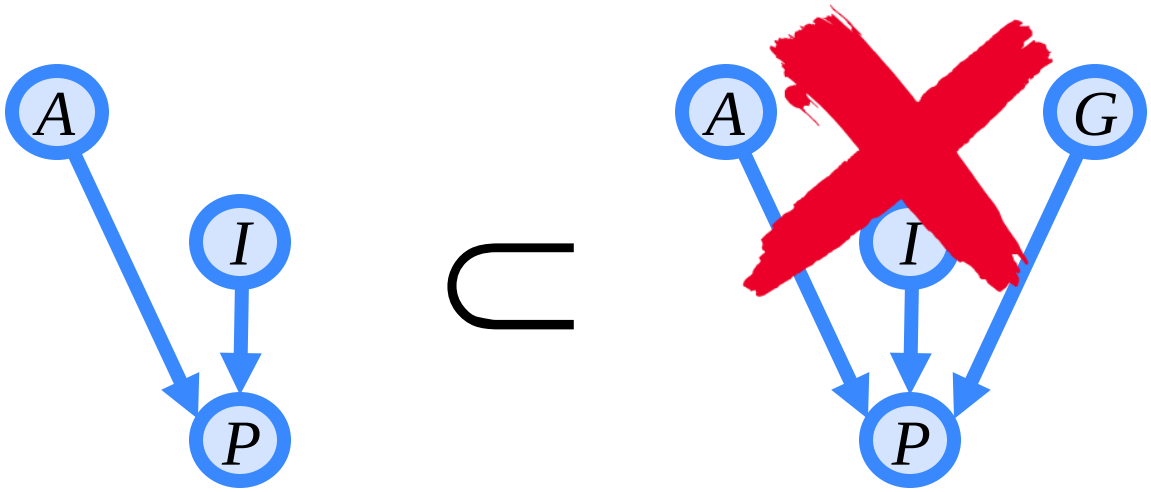
Pruning Rule: Basic Idea



SCORE(P | AI) $<$ **SCORE(P | AIG)**

[Teyssier & Koller 2005]

Pruning Rule: Basic Idea

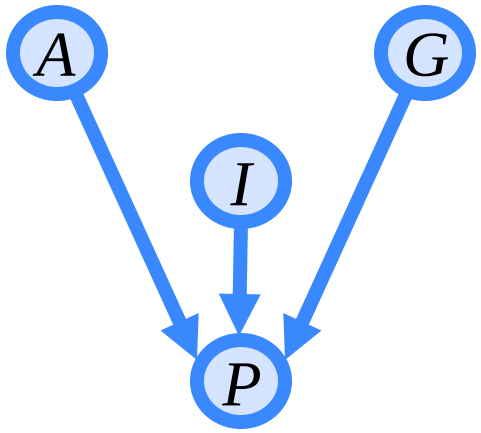


SCORE(P | AI) $<$ **SCORE(P | AIG)**

Can we generalize to the problem of enumerating the k-best BNs?

New Pruning Rule: Intuitive Idea 1

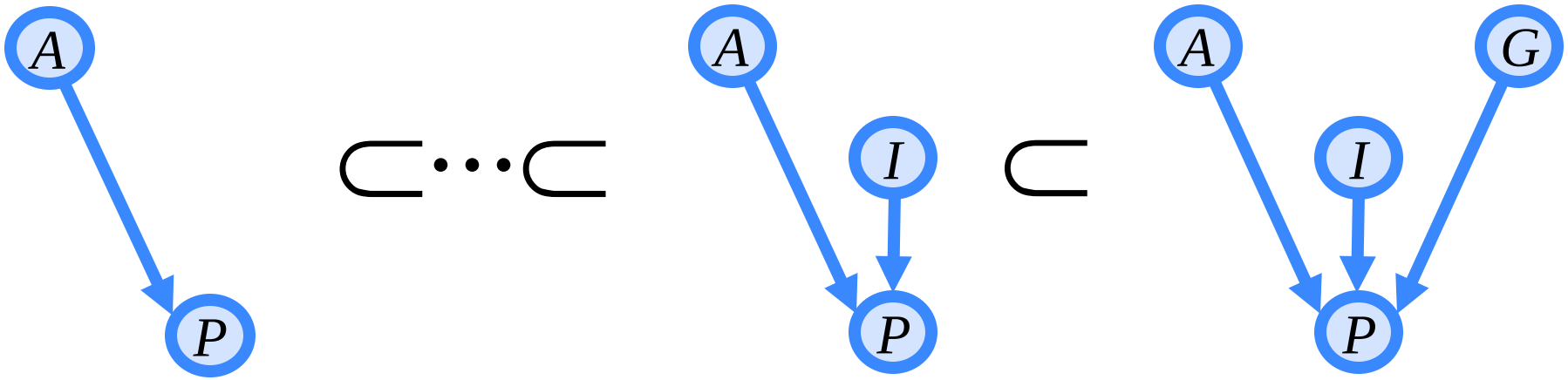
Theorem 1 [Local Test]:



SCORE(X|U)

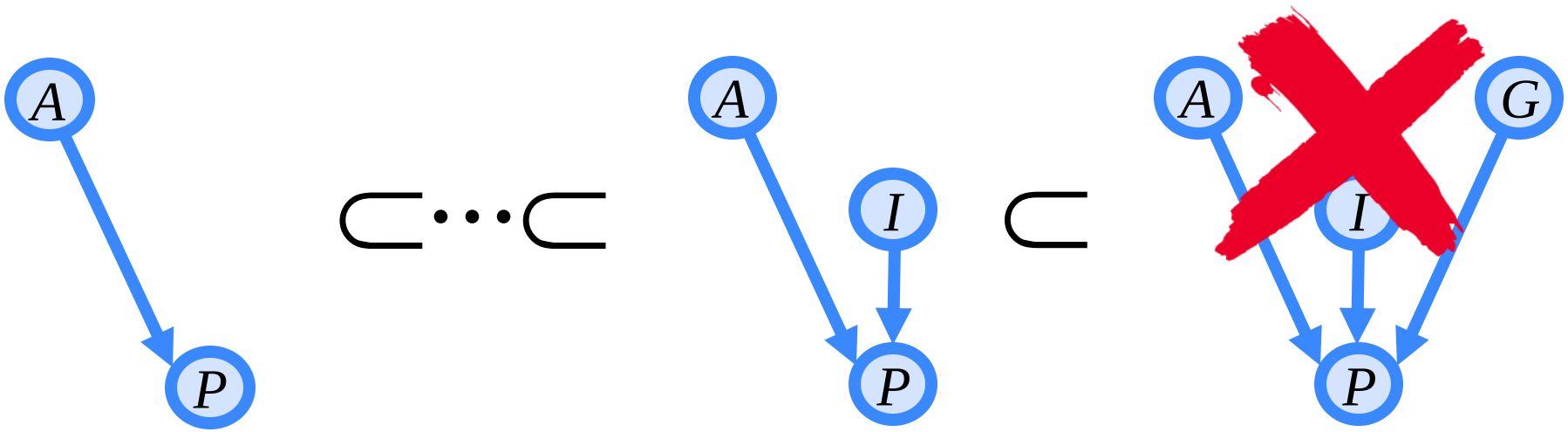
New Pruning Rule: Intuitive Idea 1

Theorem 1 [Local Test]:



New Pruning Rule: Intuitive Idea 1

Theorem 1 [Local Test]:



SCORE(X|U_1) < • • • < SCORE(X|U_k) < SCORE(X|U)

New Pruning Rule 1

- **Theorem 2:** Under the MDL score, if:

$$H_{\max}(X) \leq \frac{1}{4} \cdot \log_2 N \cdot K(X|\mathbf{U})$$

then ***every*** subset of **U** has a better score.

New Pruning Rule 1

- **Theorem 3:** Under the MDL score, do not consider families *X***U** where:

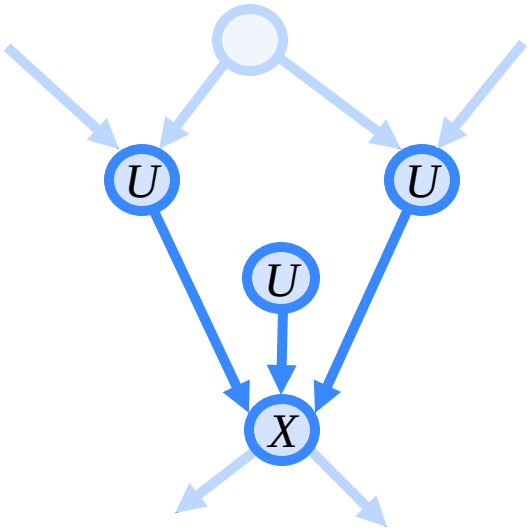
$$|\mathbf{U}| > \left\lfloor \log_2 \frac{4N}{\log_2 N} \right\rfloor$$

if there are at least *k* subsets of **U**.

Generalization of [Suzuki 1996, Tian 2000, De Campos and Ji 2011]

New Pruning Rule: Intuitive Idea 2

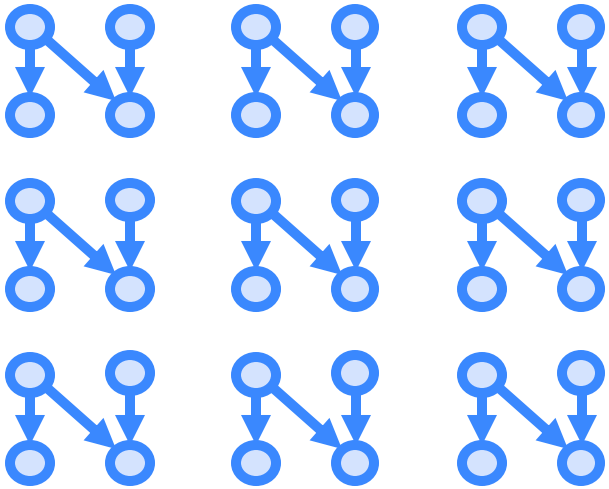
Theorem 4 [Global Test]:

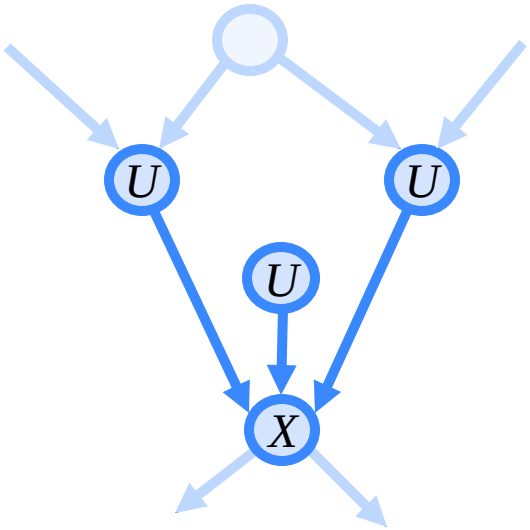


Best SCORE(G) with family $X|\mathbf{U}$

New Pruning Rule: Intuitive Idea 2

Theorem 4 [Global Test]:

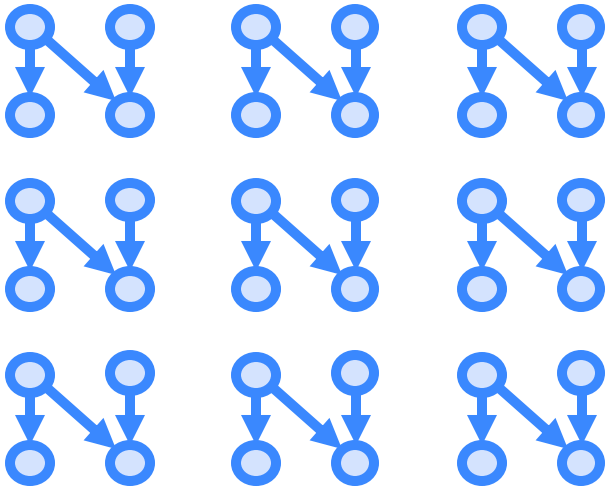


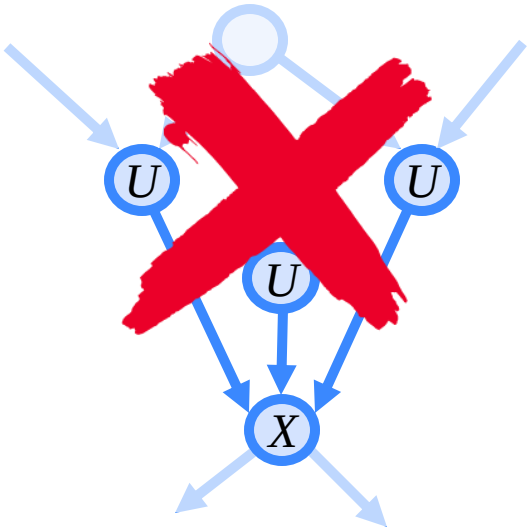


SCORE(G_1) < • • • < SCORE(G_k) < Best SCORE(G) with family X|U

New Pruning Rule: Intuitive Idea 2

Theorem 4 [Global Test]:





SCORE(G_1) < • • • < SCORE(G_k) < Best SCORE(G) with family $X|\mathbf{U}$

(DAGs can't contain family $X|\mathbf{U}$)

New Pruning Rule 2

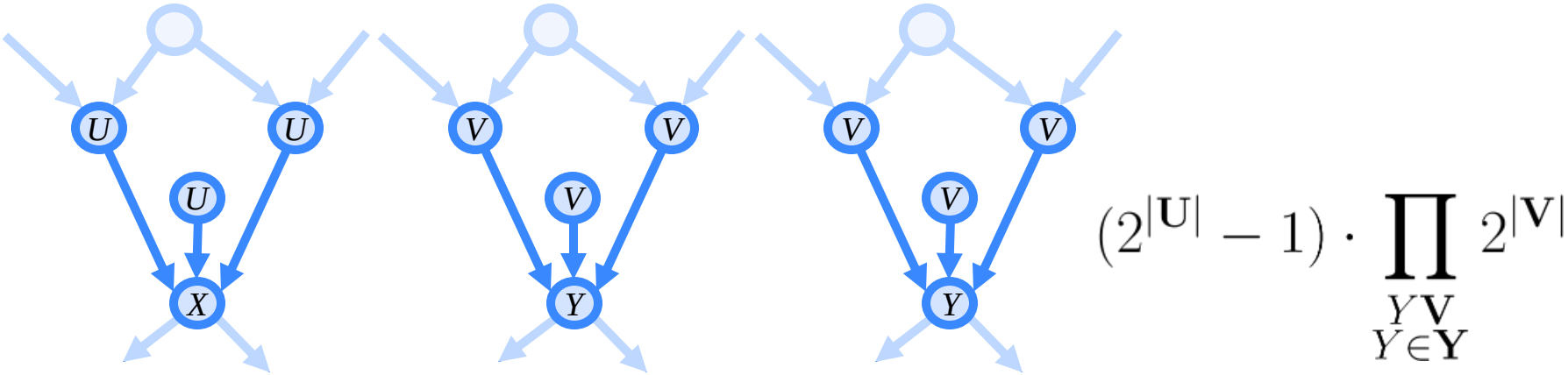
- **Theorem 5:** Under the MDL score, if a DAG *G* has families *X***U** and *Y***V**, and if:

$$H_{\max}(X) + \sum_{Y \in \mathbf{Y}} H_{\max}(Y) \leq \frac{1}{4} \cdot \log_2 N \cdot K(X|\mathbf{U})$$

then ***every*** sub-DAG of *G* w.r.t *X***U** and *Y***V** has a better score.

New Pruning Rule 2

Can find ***exponentially*** many better sub-DAGs!



$$(2^{|\mathbf{U}|} - 1) \cdot \prod_{\substack{Y \mathbf{V} \\ Y \in \mathbf{Y}}} 2^{|\mathbf{V}|}$$

Heuristic: search for DAG *G* and small set **Y** that satisfies bound for *k* we want to enumerate

Experiments

benchmark				10-best		100-best		1,000-best	
name	n	N	S	p	s	p	s	p	s
hepatitis	20	126	0.16	6	0.01	6	0.01	7	0.03
imports	22	205	0.69	6	0.03	6	0.03	7	0.07
parkinsons	23	195	1.44	6	0.04	6	0.04	8	0.10
sensors	25	5456	6.25	10	1.69	10	1.69	10	1.69
autos	26	159	13.00	6	0.10	6	0.10	8	1.46
horse	28	300	56.00	7	0.53	7	0.53	8	0.70
flag	29	194	116.00	6	0.22	6	0.22	7	0.73

n variables, *N* instances

Full score list size (*S*) vs pruned score list size (*s*) /w upper bound (*p*): orders-of-magnitude memory savings (in GB)

Experiments

benchmark		10-best			100-best			1,000-best		
name	n	E_h	T_h	T_{A*}	E_h	T_h	T_{A*}	E_h	T_h	T_{A*}
hepatitis	20	155	1.71	0.17	2188	3.32	0.80	6427	5.13	14.23
imports	22	111	63.26	0.16	232	73.83	0.20	1041	134.97	0.72
parkinsons	23	110	666.23	1.23	741	973.44	1.71	4313	3143.19	10.61
sensors	25	354	10219.25	3.65	482	13991.11	4.76	1342	23237.06	10.49
autos	26	1199	2098.97	6.46	2909	3242.36	8.96	9185	4062.17	13.78
horse	28	1095	2045.58	8.96	11653	2449.30	21.92	48069	5908.90	55.98
flag	29	1248	4454.21	19.79	26766	11093.91	45.22	110272	21959.47	257.27

[**Tian, He & Ram 2010**] enumerated 100-best for 17 variables

[**Chen, Choi & Darwiche 2015**] enumerated 1,000 best for 23 variables

Total running time is $T_h + T_{A*}$

Conclusion

- We generalized MDL pruning rules to the problem of enumerating the k-best BNs
  - local test: find *k* better families
  - global test: find *k* better DAGs
- Scale from 23 variable (no pruning) to 29 variables (with pruning)

Thanks!

Summary of Pruning Conditions

- **All subsets** of **U** are better when:

$$H_{\max}(X) \leq \frac{1}{4} \cdot \log_2 N \cdot K(X|\mathbf{U})$$

- **All sub-DAGs** w.r.t. X**U** & Y**V** are better when**:**

$$H_{\max}(X) + \sum_{Y \in \mathbf{Y}} H_{\max}(Y) \leq \frac{1}{4} \cdot \log_2 N \cdot K(X|\mathbf{U})$$