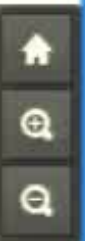


Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb



2015 SIAM International
Conference on **DATA MINING**



Scaling log-linear analysis to
datasets with 1,000+ variables

François Petitjean and Geoff Webb



2015 SIAM International
Conference on **DATA MINING**



Motivation



Log-linear analysis = one of THE standard methods in statistics

The collage contains several items:

- A document snippet with the heading "Log-linear analysis" and some text.
- A book cover with a blue and white design.
- A document snippet with a table of data.
- A software interface for "StatSoft".
- A software interface for "SSIS".
- A software interface for "Statistica".
- A document snippet with the heading "Log-linear analysis" and some text.

Log-linear analysis = one of THE standard methods in statistics

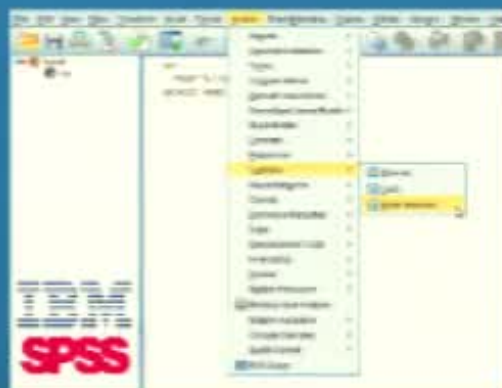
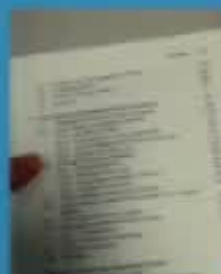


8. Loglinear Models for Contingency Tables	314
8.1 Loglinear Models for Two-Way Tables	314
8.2 Loglinear Models for Independence and Interaction in Two-Way Tables	315
8.3 Evidence for Loglinear Models	321
8.4 Loglinear Models for Higher Dimensions	326
8.5 The Loglinear-Logit Model Connection	330
8.6 Loglinear Model Fitting: Conditional Probabilities and Asymptotic Distributions	333
8.7 Loglinear Model Fitting: Iterative Methods and their Applications	342
Notes	346
Problems	347

[BOOK] **Categorical data analysis**

A Agresti - 2013 - books.google.com

Praise for the Second Edition "A must-have book for applications in **categorical data analysis**."—Statist
this book."—Pharmaceutical Research" If you do an
Cited by 17177 Related articles All 27 versions



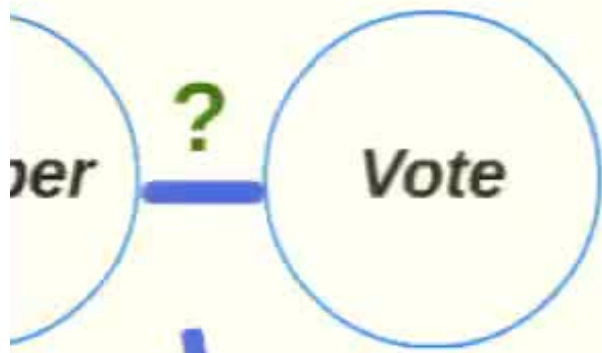
Linear analysis

...a very simple example

ing for the presidential election
ident of being a **member of an**
ation?"



. formulate hypotheses



3. Evaluation

Stats

The null hypothesis: H_0

The hypothesis: H_1

Expectation	Observed
Male	Female
Male	Male
Female	Female
Female	Male

Let's do the maths

$\chi^2 = \sum \frac{(O - E)^2}{E}$

Information theory

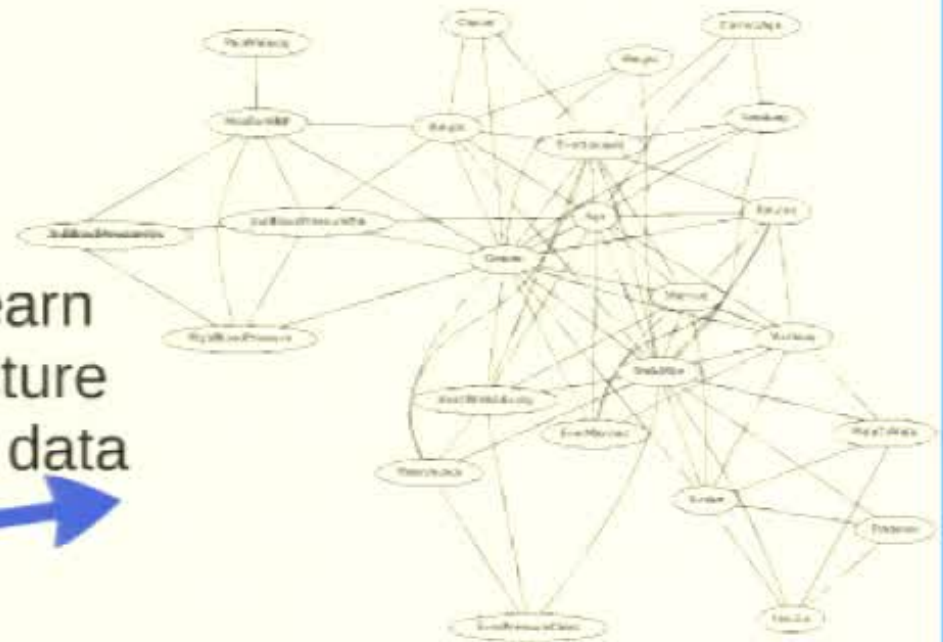
Category	Value
Male	Female
Male	Male
Female	Female
Female	Male

Log-linear analysis

Learning graphical models from data

	A	B	E	GT	GU	GV	GW	
1	Gender	Age	Working	Retire	ke	Cancer	Diabetes	Insulin
2	Male	30-39	No	Yes	No	No	No	No
3	Female	30-39	No	No	No	No	No	No
4	Male	30-39	Yes	No	Yes	No	No	No
5	Male	30-39	No	Yes	No	No	No	No
6	Female	30-39	No	No	No	No	No	No
7	Female	30-39	No	Yes	No	No	No	No
8	Female	30-39	No	No	No	No	No	No
9	Male	30-39	Yes	Yes	No	No	No	No
10	Female	30-39	No	Yes	No	No	No	No
11	Male	30-39	No	Yes	No	No	No	No
12	Female	30-39	No	No	Yes	No	No	No
13	Male	30-39	Yes	Yes	No	No	No	No
14	Female	30-39	No	Yes	No	No	No	No
15	Male	30-39	Yes	Yes	No	No	No	No
16	Female	30-39	No	Yes	No	No	No	No
17	Male	30-39	Yes	Yes	No	No	No	No
18	Male	30-39	Yes	No	Suspect	No	Suspect	No
19	Male	30-39	No	Yes	No	No	No	No
20	Female	30-39	Yes	Yes	No	No	No	No
21	Female	30-39	No	Yes	No	No	No	No
22	Male	30-39	No	Yes	Yes	No	No	No
23	Female	30-39	Yes	Yes	No	No	No	No
24	Female	30-39	No	Yes	No	No	No	No
25	Female	30-39	No	Yes	No	No	No	No
26	Male	30-39	No	Yes	Suspect	No	Suspect	No
27	Female	30-39	Yes	Yes	No	No	No	No
28	Male	30-39	Yes	No	No	No	No	No

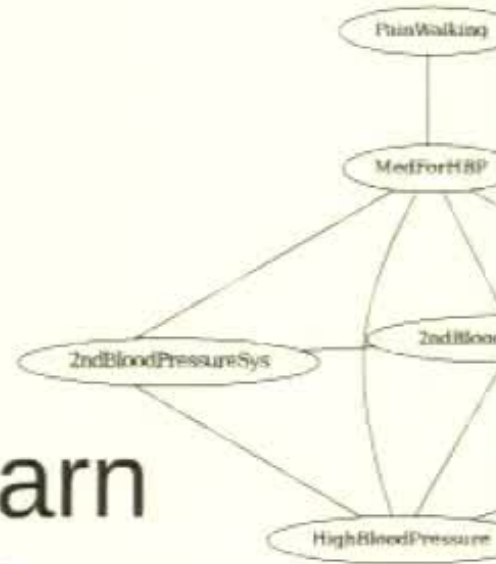
1. Learn structure from data



2. Use the structure



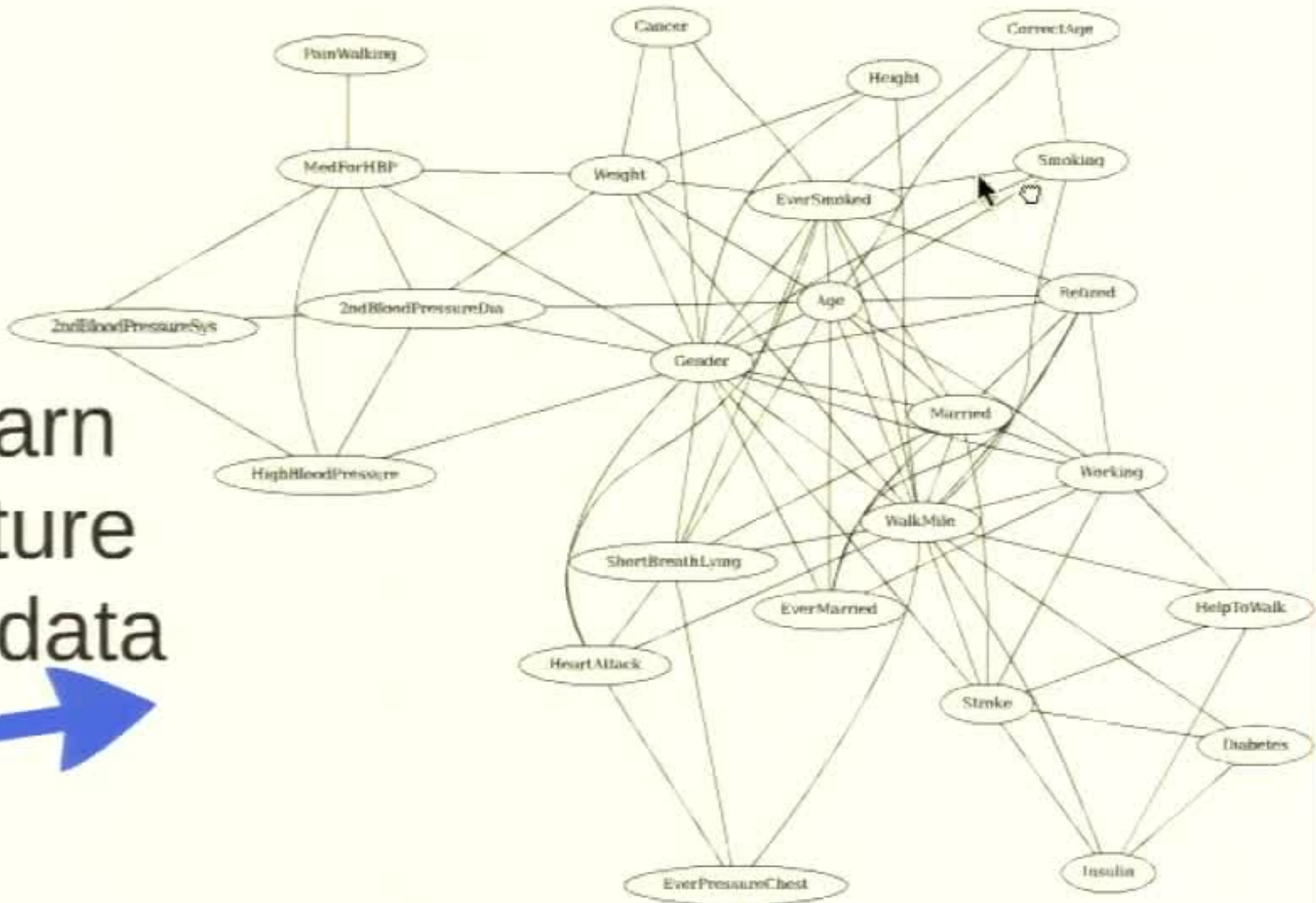
	A	B	E	F	GT	GU	GV	GW
1	Gender	Age	Working	Retire	Stroke	Cancer	Diabetes	Insulin
2	Male	85over	No	Yes	No	No	No	No
3	Female	85over	No	No	No	No	No	No
4	Male	85over	?	?	No	Yes	No	No
5	Male	80-84	No	Yes	No	No	No	No
6	Female	80-84	No	No	No	No	No	No
7	Female	85over	No	Yes	No	No	No	No
8	Female	80-84	No	No	No	No	No	No
9	Male	80-84	No	Yes	No	No	No	No
10	Female	80-84	No	Yes	No	No	No	No
11	Male	80-84	No	Yes	No	No	No	No
12	Female	75-79	No	No	No	Yes	No	No
13	Male	80-84	Yes	Yes	No	No	No	No
14	Female	80-84	No	Yes	No	No	No	No
15	Male	75-79	No	Yes	No	No	No	No
16	Female	80-84	No	Yes	No	No	No	No
17	Male	75-79	No	Yes	No	No	No	No
18	Male	75-79	Yes	No	Suspect	No	Suspect	No
19	Male	80-84	No	Yes	No	No	No	No
20	Female	75-79	No	Yes	No	No	No	No
99980	Male	70-74	No	Yes	No	Yes	No	No
99981	Male	70-74	No	No	No	No	No	No
99982	Female	70-74	Yes	Yes	No	No	No	No
99983	Female	70-74	No	Yes	No	No	No	No
99984	Female	70-74	No	Yes	No	No	No	No
99985	Female	70-74	No	Yes	No	No	No	No
99986	Male	70-74	No	Yes	Suspect	No	Suspect	No
99987	Female	70-74	No	Yes	No	No	No	No
99988	Male	70-74	Yes	No	No	No	No	No



1. Learn structure from data



cal models from data

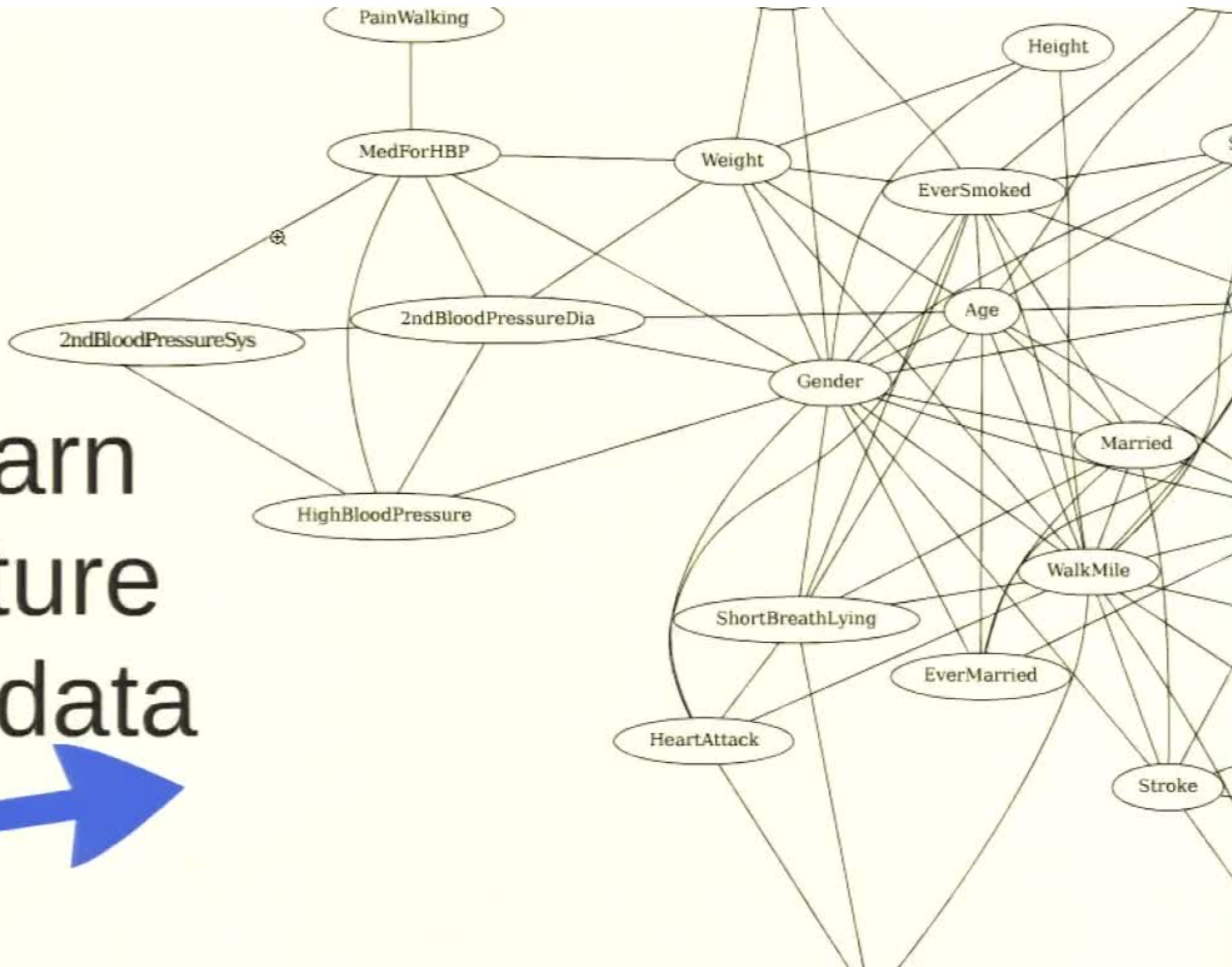


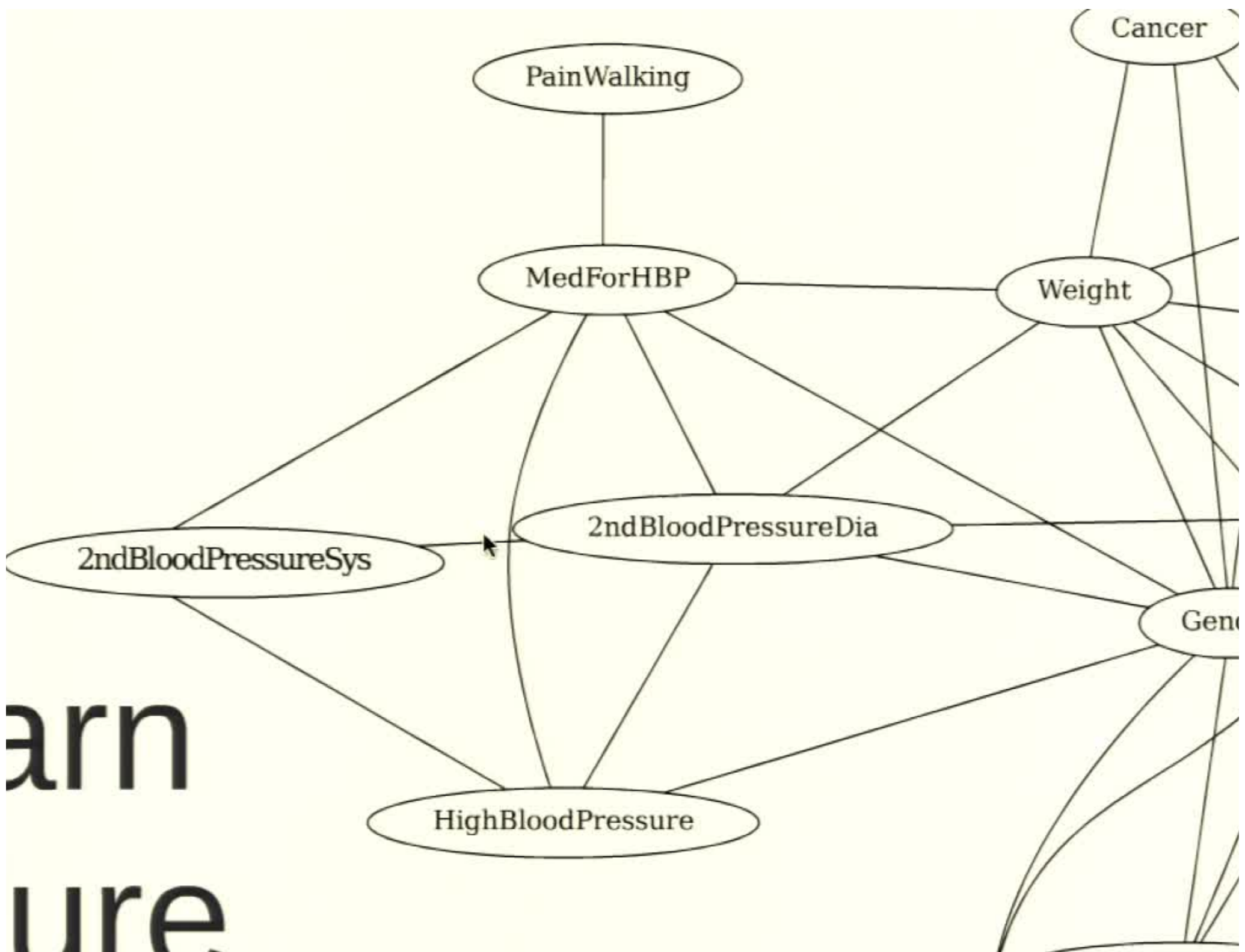
1. Learn structure from data



GW
ulin

Learn
structure
data





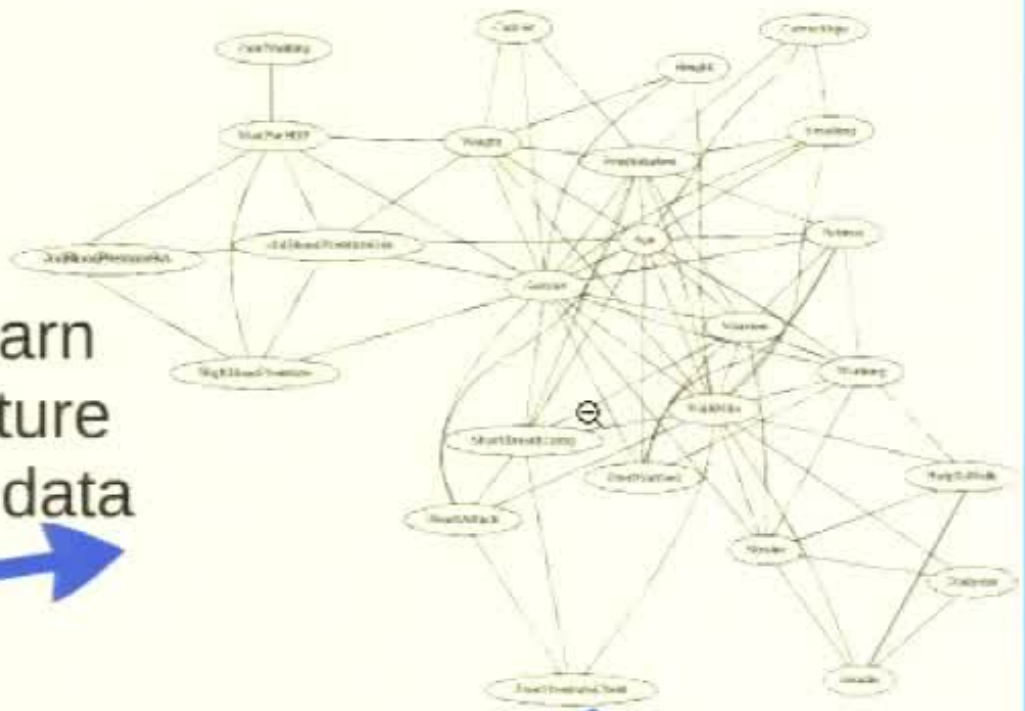
arn
ure

log-linear analysis

Learning graphical models from data

	A	B	E	F	CT	CU	CV	GW
	Gender	Age	Working	Retire	Smoke	Cancer	Diabetes	Insulin
1	Male	55over	No	Yes	No	No	No	No
2	Female	55over	No	No	No	No	No	No
3	Male	55over	?	?	Yes	No	No	No
4	Male	55-64	No	Yes	No	No	No	No
5	Female	55-64	No	No	No	No	No	No
6	Female	55-64	No	Yes	No	No	No	No
7	Female	55-64	No	Yes	No	No	No	No
8	Female	55-64	No	No	No	No	No	No
9	Male	55-64	No	Yes	No	No	No	No
10	Female	55-64	No	Yes	No	No	No	No
11	Male	55-64	No	Yes	No	No	No	No
12	Female	55-64	No	No	Yes	No	No	No
13	Male	55-64	Yes	Yes	No	No	No	No
14	Female	55-64	No	Yes	No	No	No	No
15	Male	55-64	No	Yes	No	No	No	No
16	Female	55-64	No	Yes	No	No	No	No
17	Male	55-64	No	Yes	No	No	No	No
18	Male	55-64	Yes	No	Suspect	No	Suspect	No
19	Male	55-64	No	Yes	No	No	No	No
20	Male	55-64	Yes	Yes	No	No	No	No
21	Male	55-64	No	Yes	No	No	No	No
22	Male	55-64	No	Yes	Yes	No	No	No
99981	Male	70-74	No	No	No	No	No	No
99982	Female	70-74	Yes	Yes	No	No	No	No
99983	Female	70-74	No	Yes	No	No	No	No
99984	Female	70-74	No	Yes	No	No	No	No
99985	Female	70-74	No	Yes	No	No	No	No
99986	Male	70-74	No	Yes	Suspect	No	Suspect	No
99987	Female	70-74	No	Yes	No	No	No	No
99988	Male	70-74	Yes	No	No	No	No	No

1. Learn structure from data

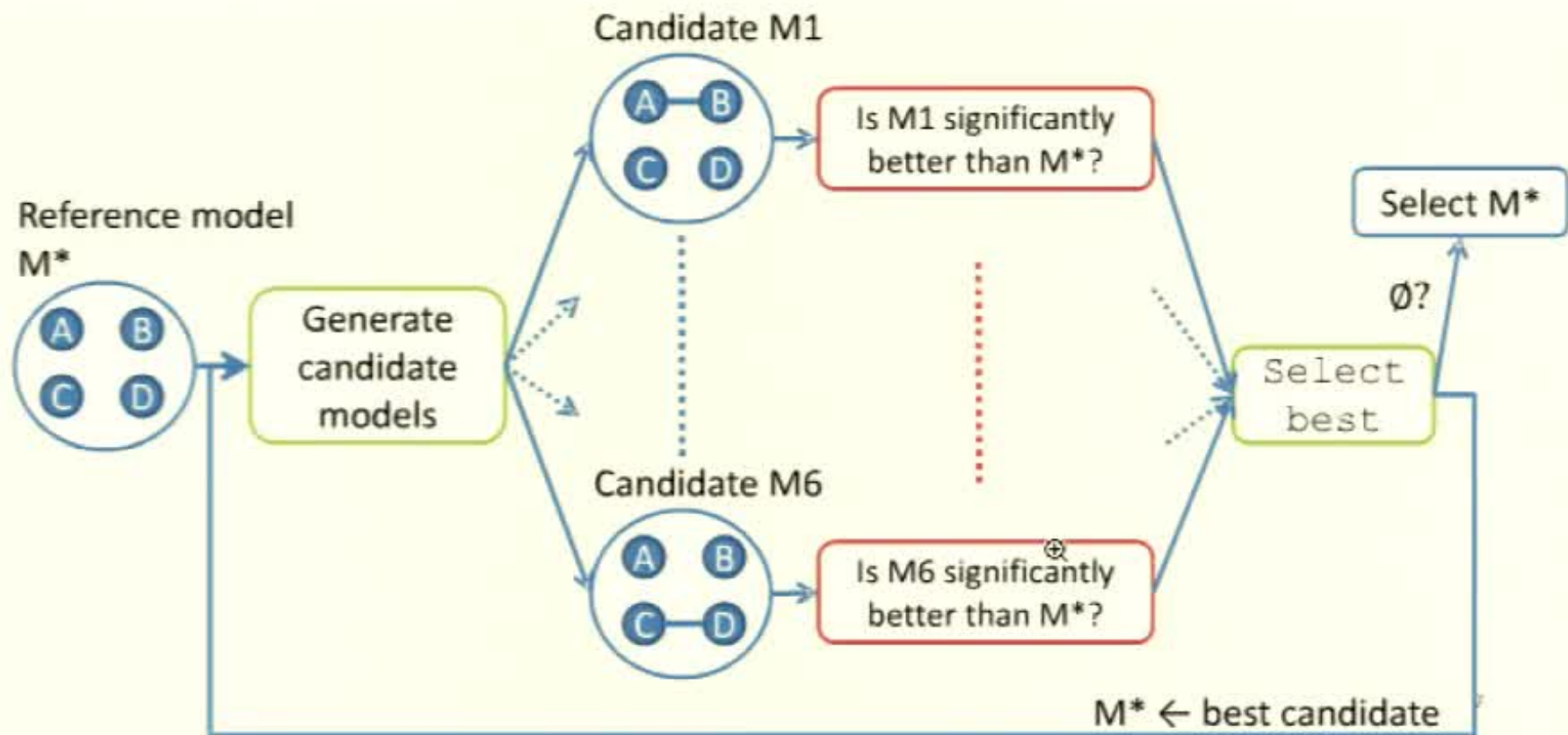


2. Use the structure



Log-linear analysis for more than 2 variables

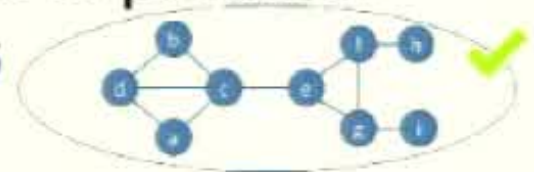
General framework: selecting a statistically significant log-linear model = superset of Markov Networks



What have shown that...

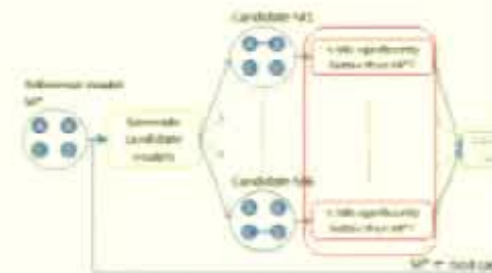
Scalability to datasets with 100 variables is possible for the class of **decomposable models**

- ICDM 2013: statistical tests
- ICDM 2014: minimum description length



Limitation: process quadratic with the number of variable

→ 1,000+ variables => **several days** of computation



What this paper shows:

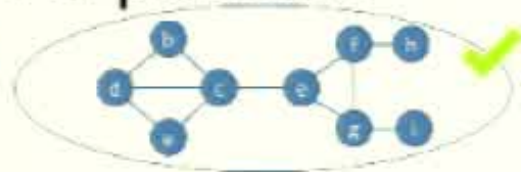
We can gain **4 orders of magnitude** while getting **exactly** the same results.

→ 1,000+ variables => **1 minute** of computation

What have shown that...

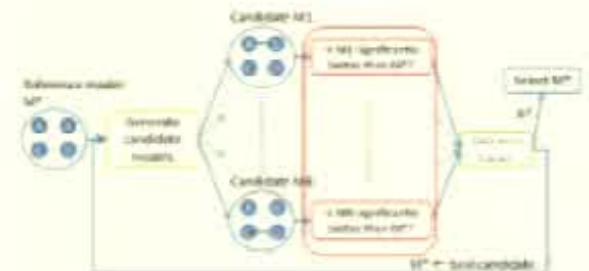
Scalability to datasets with 100 variables is possible for the class of **decomposable models**

- ICDM 2013: statistical tests
- ICDM 2014: minimum description length



Limitation: process quadratic with the number of variables

→ 1,000+ variables => **several days** of computation



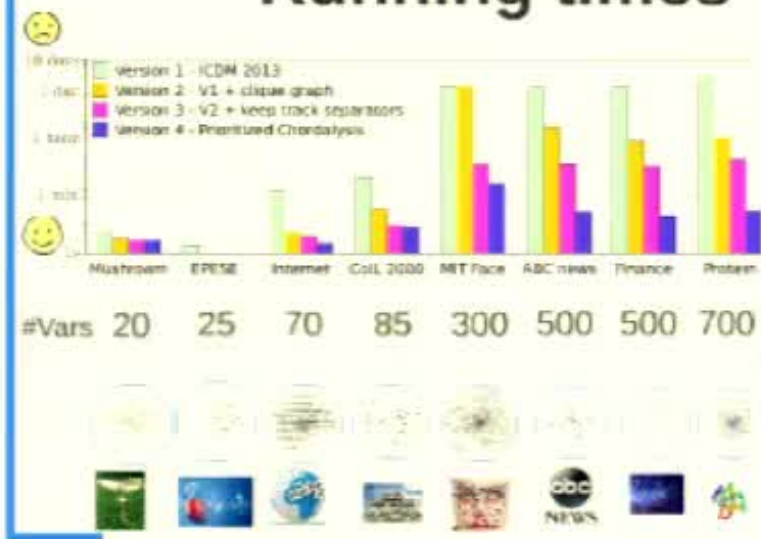
What this paper shows:

*We can gain **4 orders of magnitude** while getting **exactly** the same results.*

→ 1,000+ variables => **1 minute** of computation

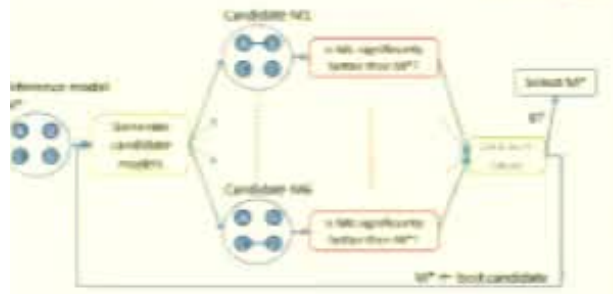
zed lysis

Running times



g-linear analysis for more than 2 variables

General framework: selecting a statistically significant linear model + superclass of Markov Networks



What have shown that...

Scalability to datasets with 100 variables is possible for the class of decomposable models

- ICDM 2013: statistical tests
- ICDM 2014: minimum description length



Limitation: process quadratic with the number of variables

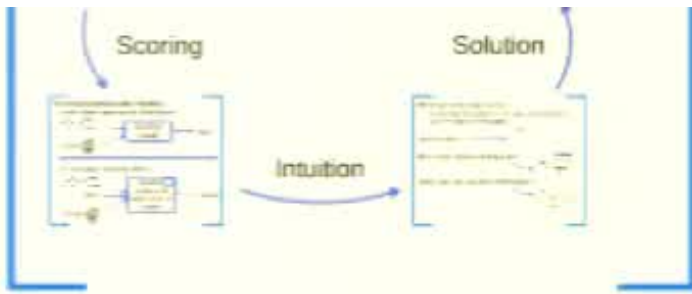
→ 1,000+ variables ⇒ **several days** of computation



What this paper shows:

We can gain **4 orders of magnitude** while getting exactly the same results.

→ 1,000+ variables ⇒ **1 minute** of computation

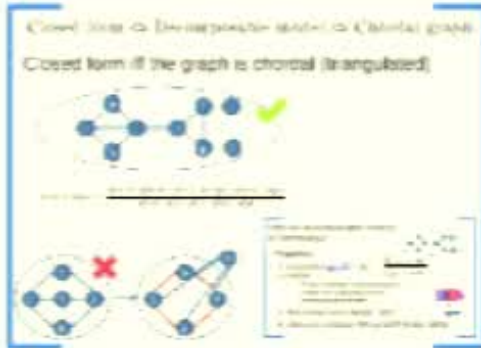


It wo

Prioritized Chordalysis

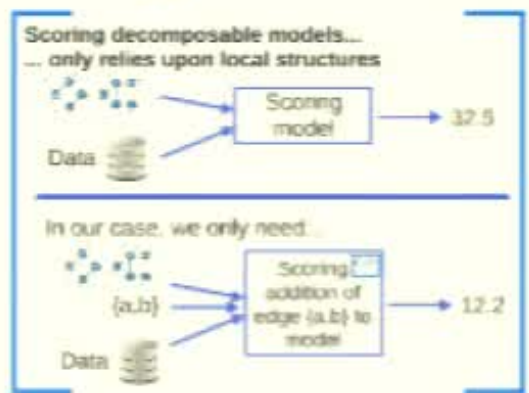


Prioritized Chordality

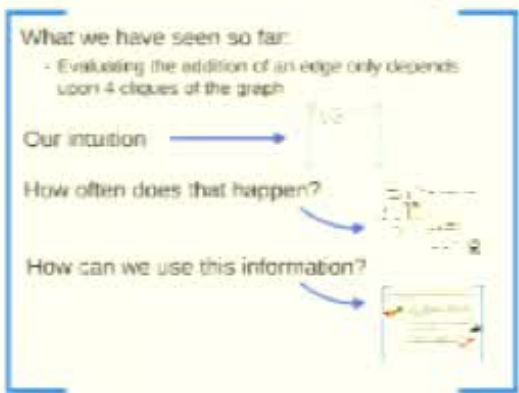


Scoring

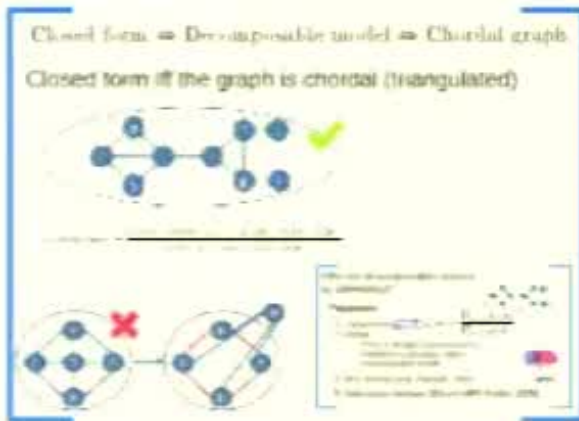
Solution



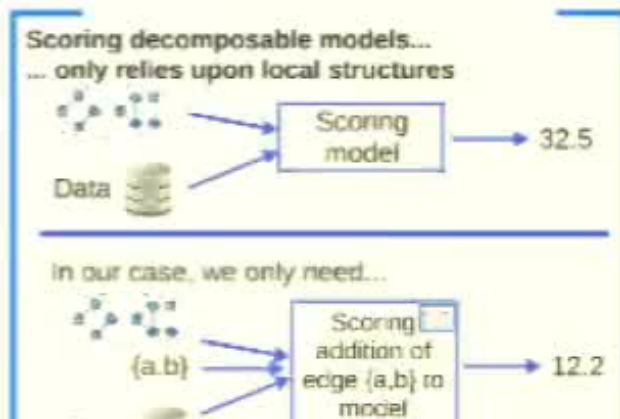
Intuition



Prioritized Chordalys



Scoring



Intuition

Solution

What we have seen so far:

- Evaluating the addition of an edge on upon 4 cliques of the graph

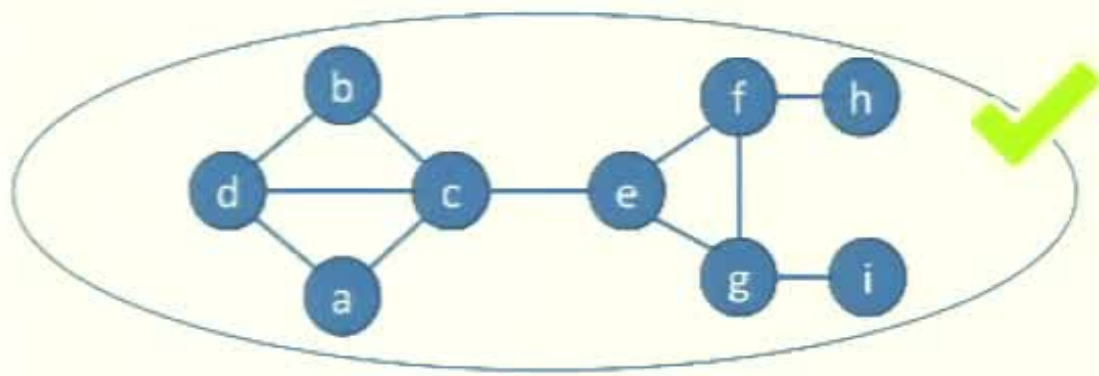
Our intuition \rightarrow

How often does that happen?

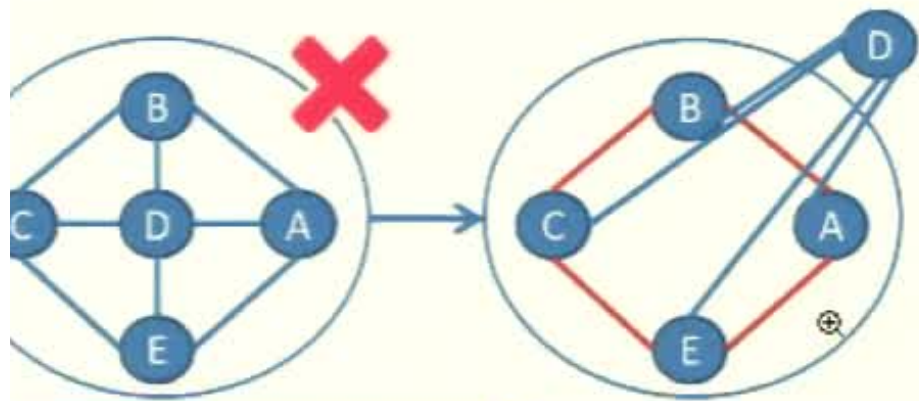
How can we use this information?

Closed form \Leftrightarrow Decomposable model \Leftrightarrow Chordal graph

Closed form iff the graph is chordal (triangulated)



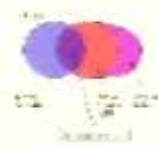
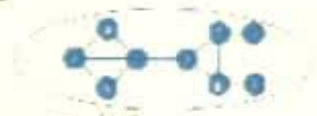
$$p(abcde fghi) = \frac{p(acd) \cdot p(bcd) \cdot p(ce) \cdot p(efg) \cdot p(fh) \cdot p(gi)}{p(cd) \cdot p(c) \cdot p(e) \cdot p(f) \cdot p(g)}$$



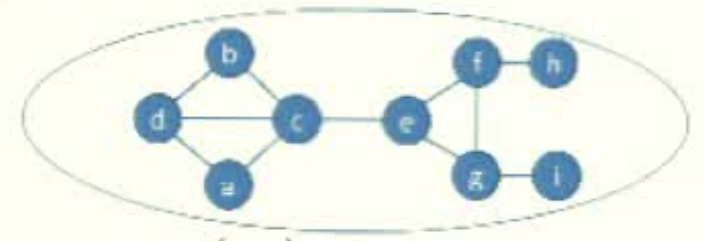
Why are decomposable models so interesting?

Properties:

1. Closed form $\Leftrightarrow p_{\mu}(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} p_C(\mathbf{x})}{\prod_{S \in \mathcal{S}} p_S(\mathbf{x})}$
2. Useful:
 - There is always a decomposable model that subsumes a non-decomposable model
3. MLE always exist [Agresti, 2002]
4. Intersection between BN and MRF [Koller, 2009]



Why are decomposable models so *interesting*?



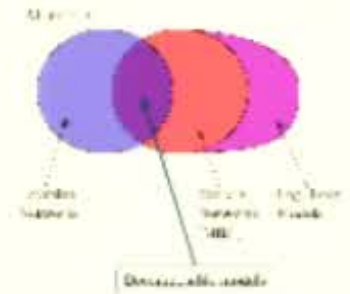
Properties:

1. Closed form $\iff p_{\mu}(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} p_C(\mathbf{x})}{\prod_{S \in \mathcal{S}} p_S(\mathbf{x})}$
2. Useful:

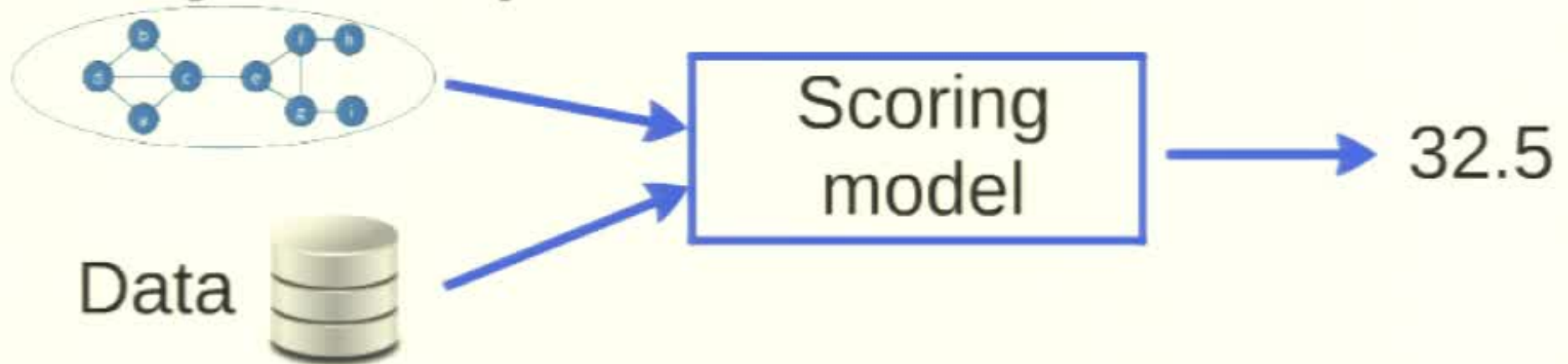
- There is always a decomposable model that subsumes a non-decomposable model

3. MLE always exist [Agresti, 2002]

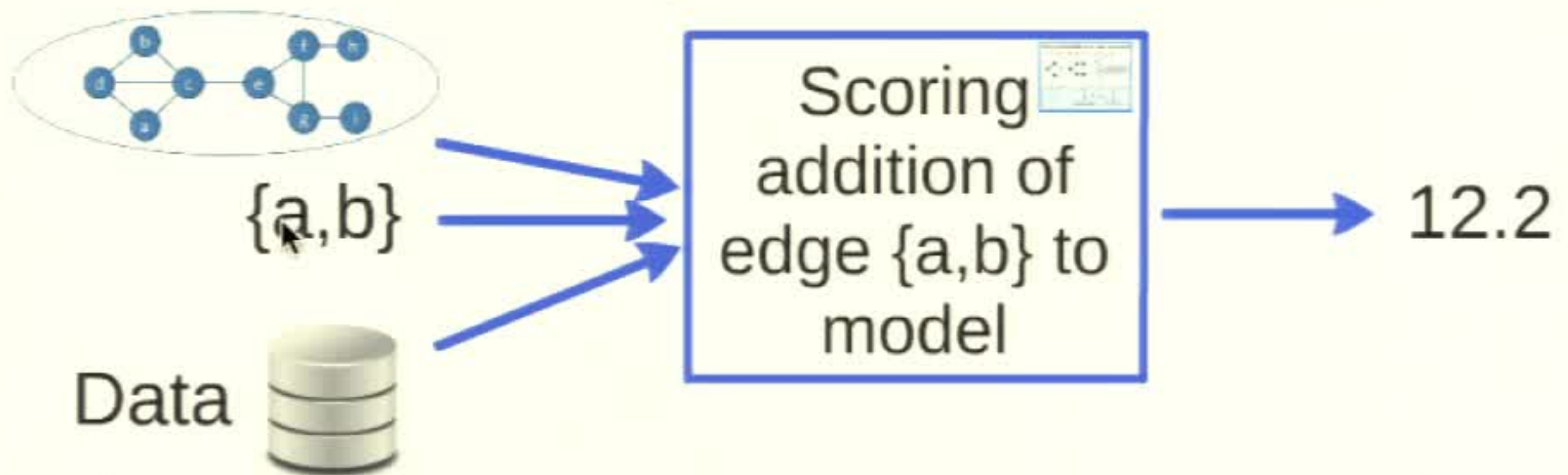
4. Intersection between BN and MRF [Koller, 2009]



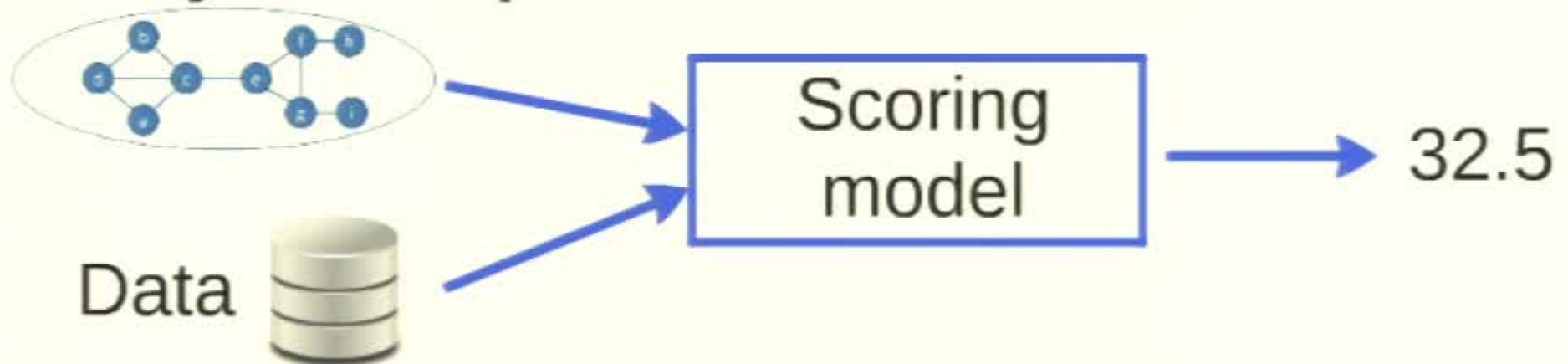
Scoring decomposable models... ... only relies upon local structures



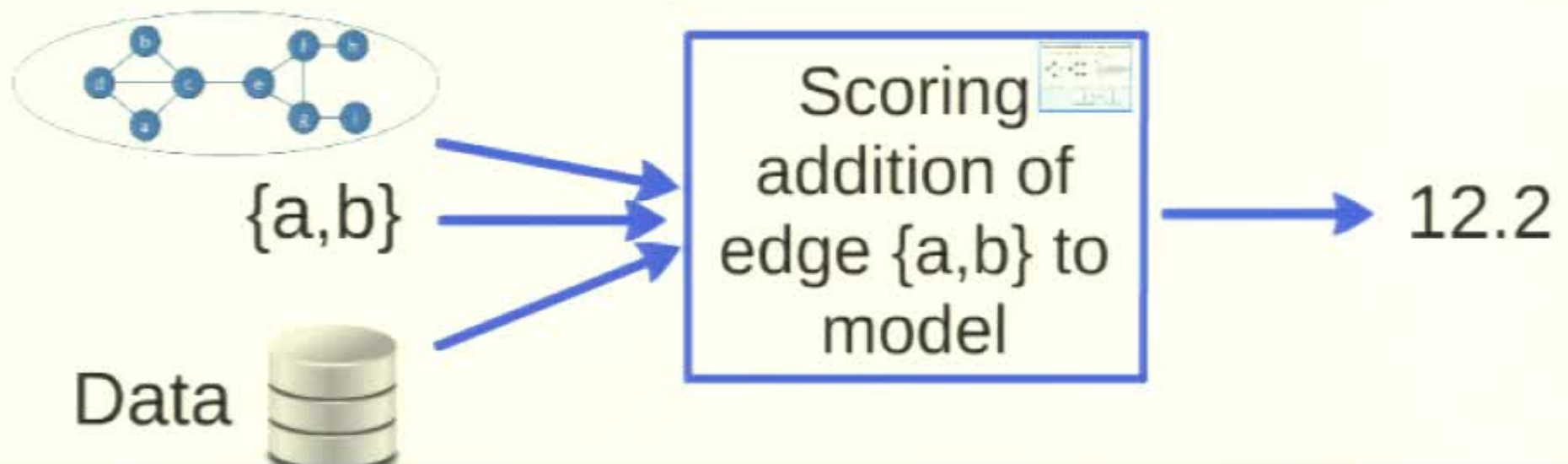
In our case, we only need...



Scoring decomposable models... ... only relies upon local structures



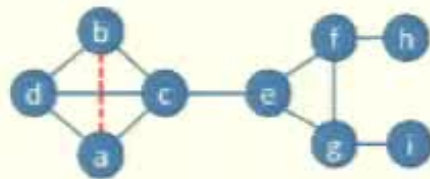
In our case, we only need...



ing

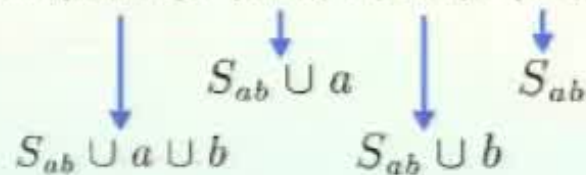
Scoring the addition of an edge to a model

$$\text{score}(\mathcal{M}, (a, b), \mathcal{D}) = \text{score}'(a, b, S_{ab}, \mathcal{D})$$



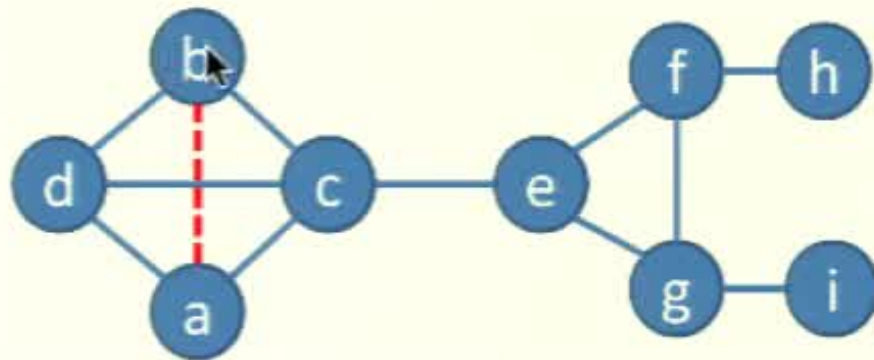
S_{ab} : minimal separator of (a,b)
= minimal set of vertices that would disconnect a from b if removed from the graph
= {c,d}

$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$



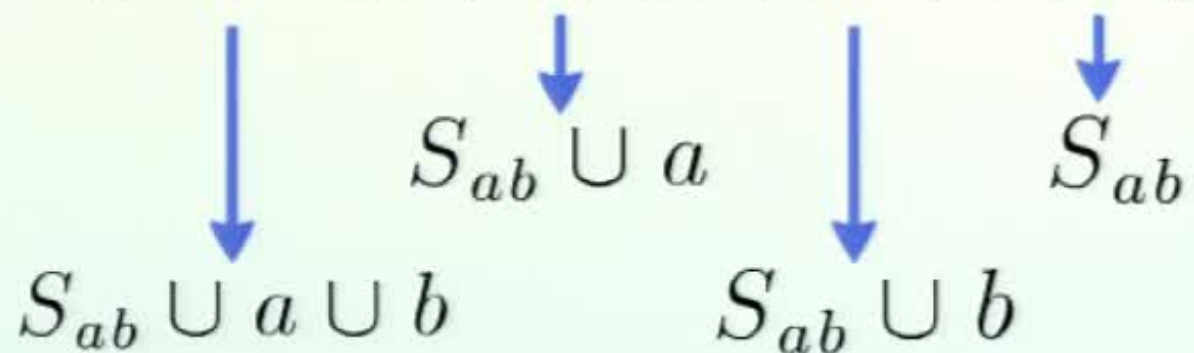
Scoring the addition of an edge to a model

$$\text{score}(\mathcal{M}, (a, b), \mathcal{D}) = \text{score}'(a, b, S_{ab}, \mathcal{D})$$

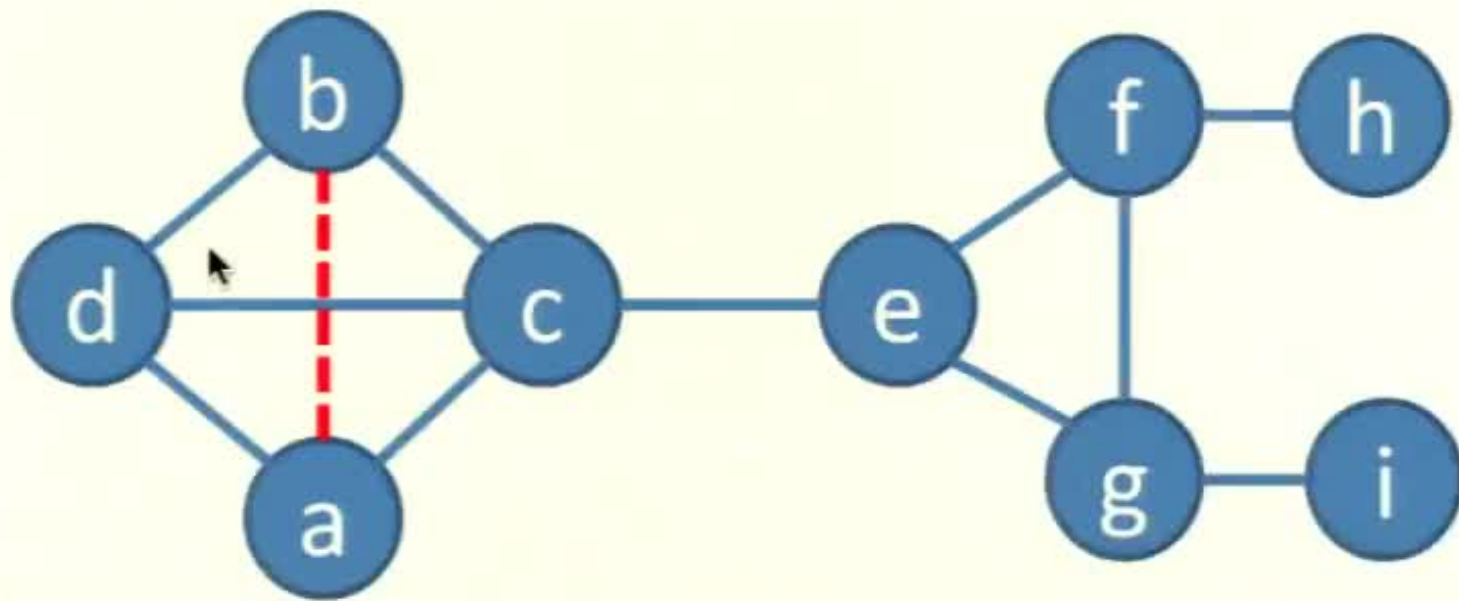


S_{ab} : minimal separator of (a,b)
 = **minimal set of vertices** that would **disconnect** a from b if removed from the graph
 = {c,d}

$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$



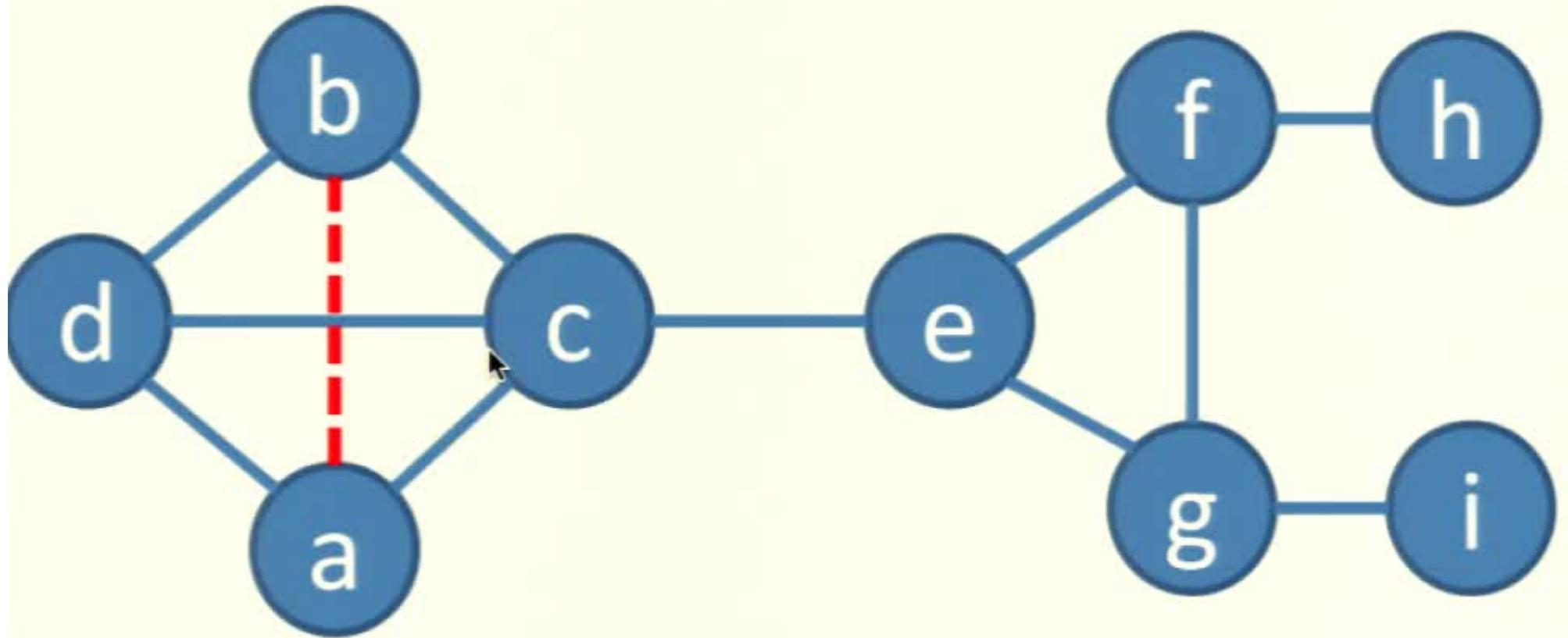
$$\text{score}(\mathcal{M}, (a, b), \mathcal{D}) = \text{score}'(a, b, \mathcal{D})$$



$S_{ab} : n$
 $=$
 wo
 re
 $=$

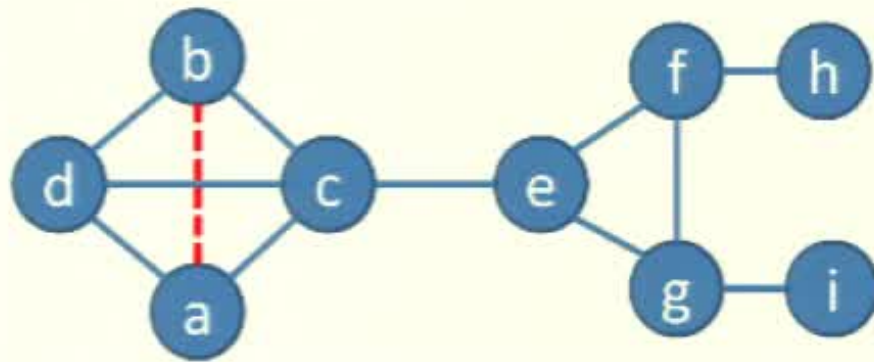
$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, b\})$$

$$\text{SCORE}(\mathcal{M}, (u, v), \mathcal{D}) = \text{SC}$$



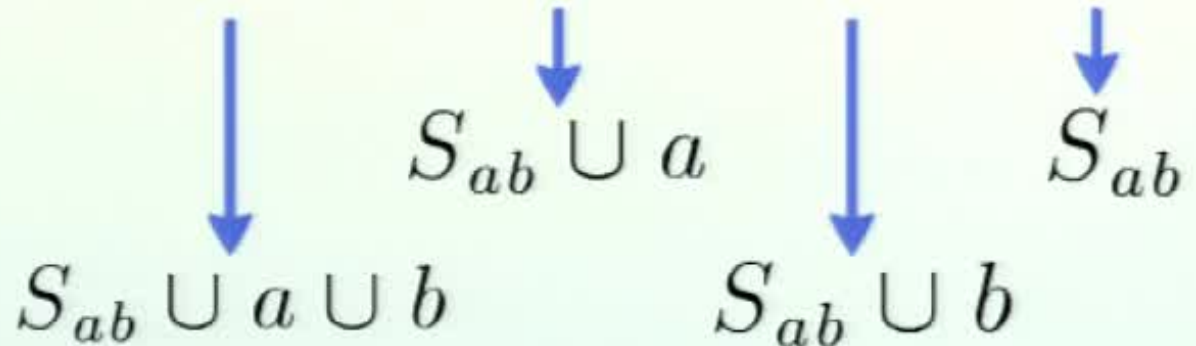
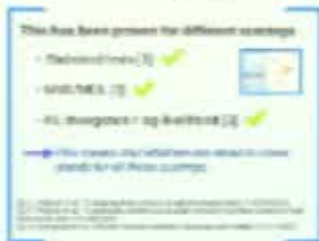
Scoring the addition of an edge to a model

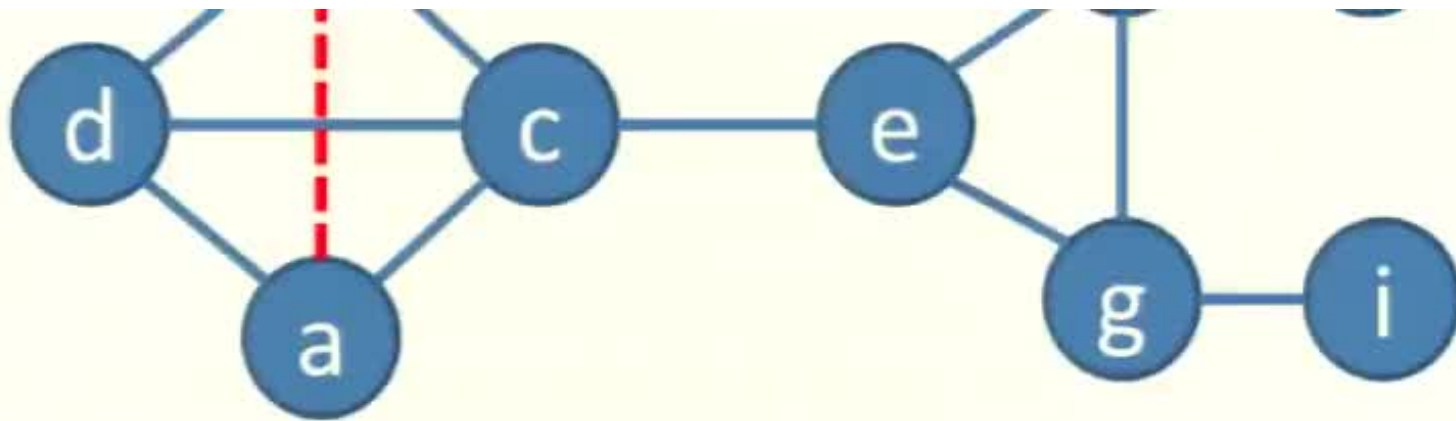
$$\text{score}(\mathcal{M}, (a, b), \mathcal{D}) = \text{score}'(a, b, S_{ab}, \mathcal{D})$$



S_{ab} : minimal separator of (a,b)
 = **minimal set of vertices** that would **disconnect** a from b if removed from the graph
 = {c,d}

$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$





$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{c, d\})$$

This has been proven for different scorings:

- Statistical tests [1] ✓
- MML/MDL [2] ✓
- KL divergence / log-likelihood [3] ✓

→ This means that what we are about to show stands for all these scorings.

[1] G. H. Golub et al., "Statistical tests for hypothesis testing in the context of causal discovery", in ICML 2014.
 [2] R. F. dechter et al., "A fast algorithm for learning the structure of Bayesian networks", in Proceedings of the AAAI Conference on Artificial Intelligence, 1994.
 [3] S. Kullback and R. A. Leibler, "On information and entropy", in IEEE Transactions on Information Theory, vol. 8, pp. 179-191, 1951.



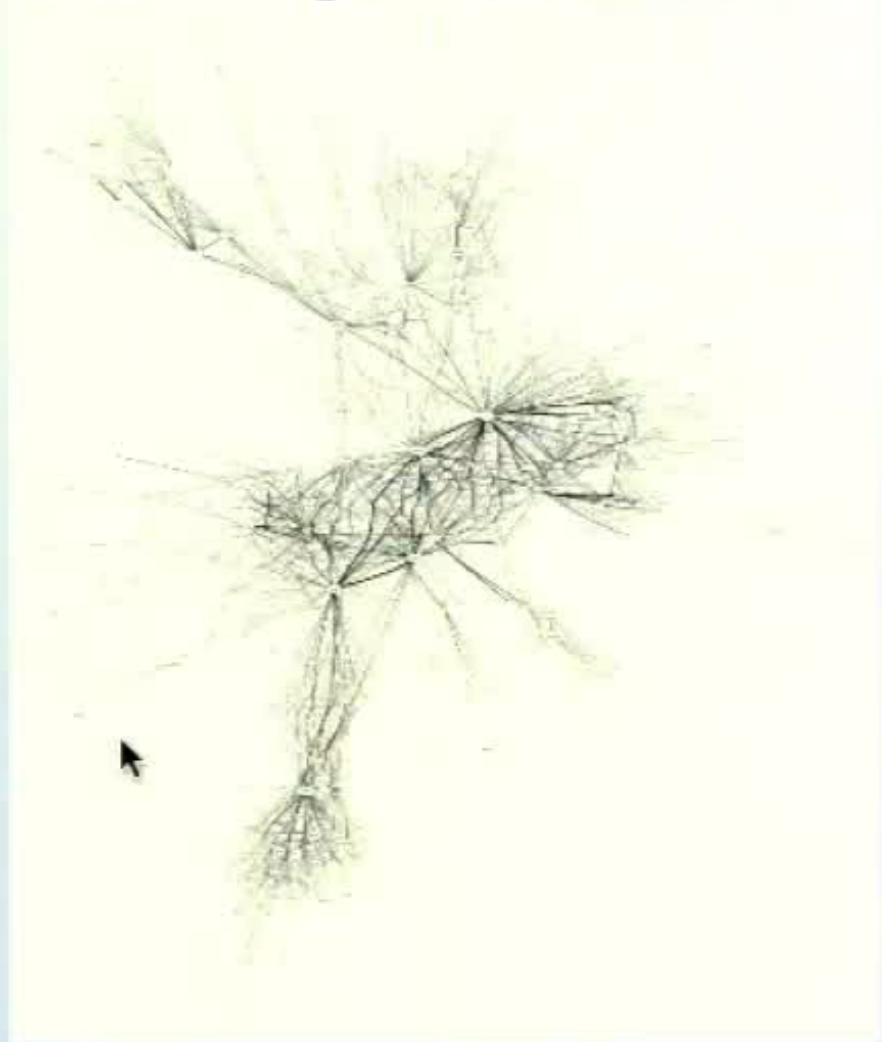
$$S_{ab} \cup a \cup b$$

S_c

Assessing the addition of one edge to this model?



We only need to consider **4 cliques**



This has been proven for different scorings

- Statistical tests [1] ✓
- MML/MDL [2] ✓
- KL divergence / log-likelihood [3] ✓




→ *This means that what we are about to show stands for all these scorings.*

[1]: F. Petitjean *et al.*, "Scaling log-linear analysis to high-dimensional data," in *ICDM 2013*.

[2]: F. Petitjean *et al.*, "A statistically efficient and scalable method for log-linear analysis of high-dimensional data," in *ICDM 2014*.

[3]: A. Deshpande *et al.*, "Efficient stepwise selection in decomposable models," in *UAI 2001*.


$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}(\mathcal{M}, \{a, b\})$$

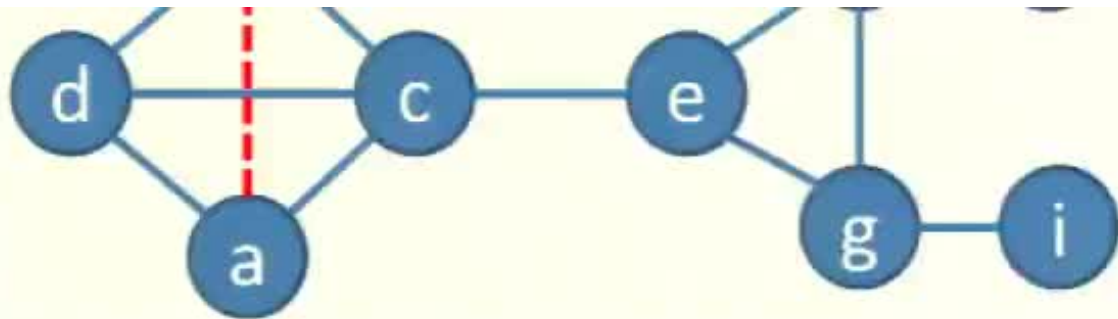
This has been proven for different scorings

- Statistical tests [1] ✓
- MML/MDL [2] ✓
- KL divergence / log-likelihood [3] ✓



→ This means that what we are about to show stands for all these scorings.

[1] F. Perron et al., "Scaling log-linear analysis to high-dimensional data," in ICDM 2013.
[2] F. Perron et al., "A statistically efficient and scalable method for log-linear analysis of high-dimensional data," in ICDM 2014.
[3] A. Deshpande et al., "Efficient stepwise selection in decomposable models," in UAI 2001.



$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\},$$

This has been proven for different scorings

- standard tests [1] ✓
- NMR ADEL [2] ✓
- VL divergence / log-likelihood [3] ✓

→ This means the references are about to show consistency of these scorings.

[1] P. Fagerberg et al., "Nucleotide sequence divergence metrics: a comparison", *Genetics*, vol. 178, no. 4, pp. 2001-2010, 2007.
[2] P. Fagerberg et al., "Nucleotide sequence divergence metrics: a comparison", *Genetics*, vol. 178, no. 4, pp. 2001-2010, 2007.
[3] P. Fagerberg et al., "Nucleotide sequence divergence metrics: a comparison", *Genetics*, vol. 178, no. 4, pp. 2001-2010, 2007.



$$S_{ab} \cup a \cup i$$

models...
structures

oring
del → 32.5

...

oring
tion of
{a,b} to
odel → 12.2

Intuition

What we have seen so far

- Evaluating the addition upon 4 cliques of the

Our intuition

How often does that

How can we use this

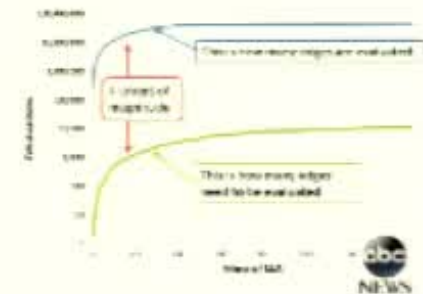
What we have seen so far:

- Evaluating the addition of an edge only depends upon 4 cliques of the graph

Our intuition



How often does that happen?



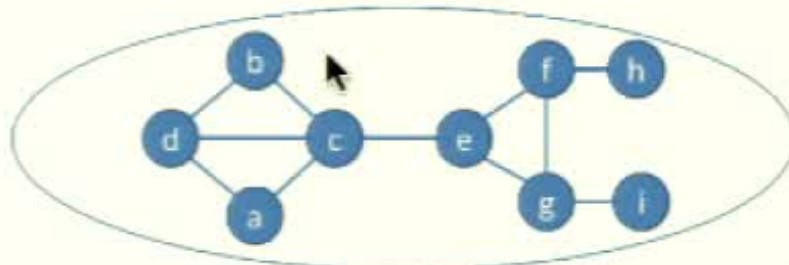
How can we use this information?



We know: $2^5 = 32$ basic change between different configurations of the graph, but the cost can be $2n$ (at least) for $n=5$ nodes.

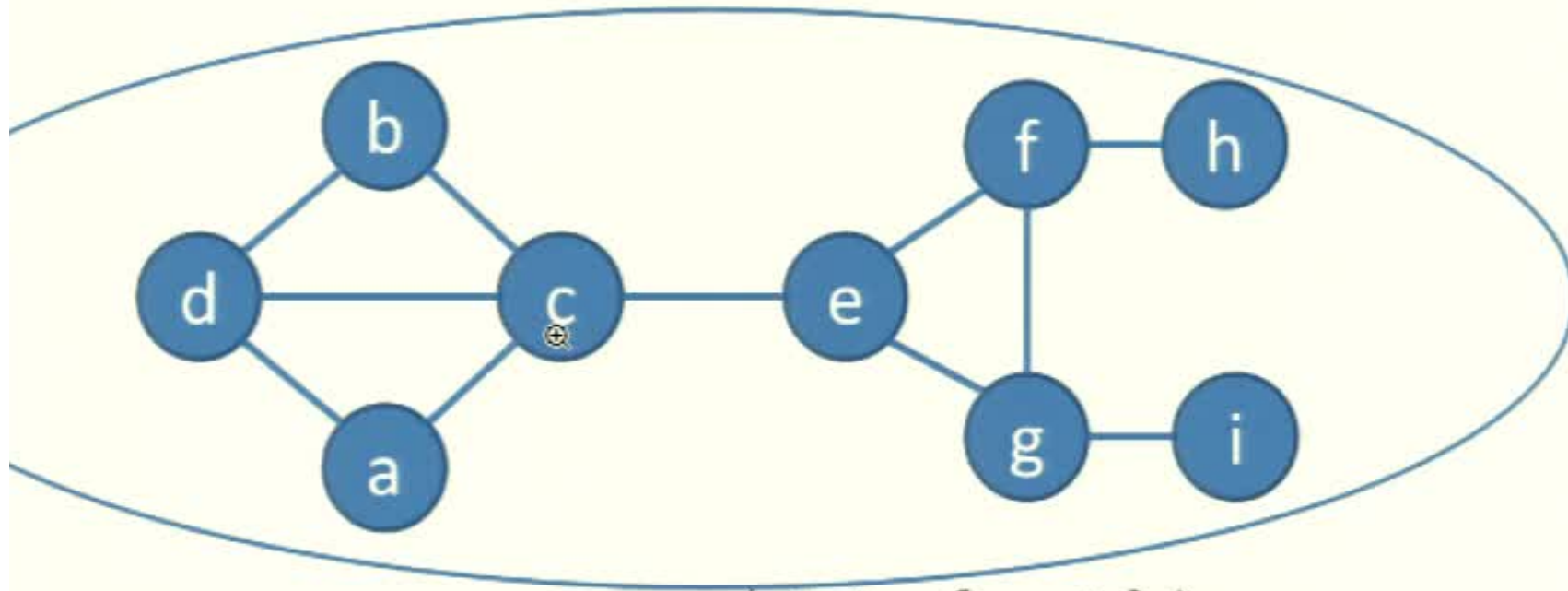
1. Use a data structure that gives direct access to inverse separators for every potential edge.
2. Keep track of the minimal separators for every potential edge.
3. Maintain an ordered list of all the potential edges (priority queue).

Not all edges should be re-examined at every step of the process



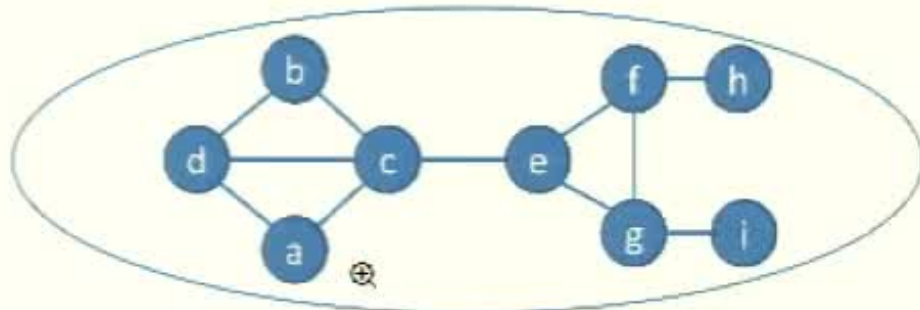
$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$

e process



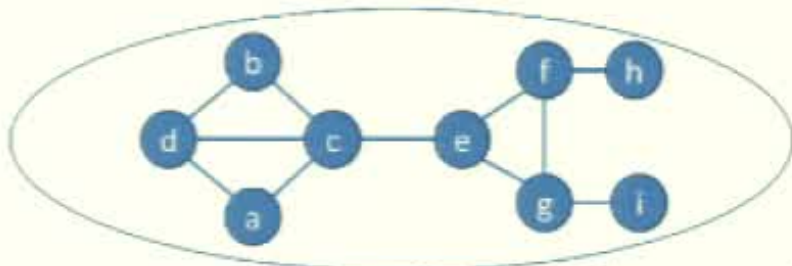
$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d, e, f, g, h, i\})$$

Not all edges should be re-examined at every step of the process



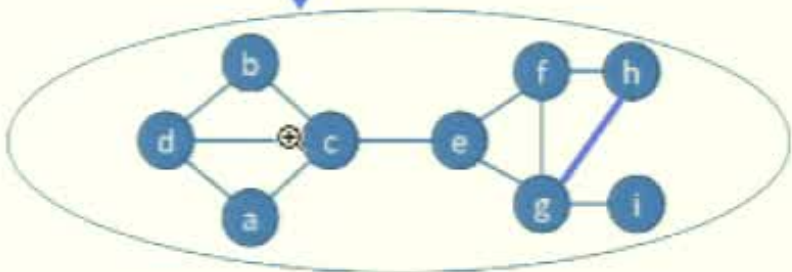
$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d$$

Not all edges should be re-examined at every step of the process



$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$

Select edge {g,h}



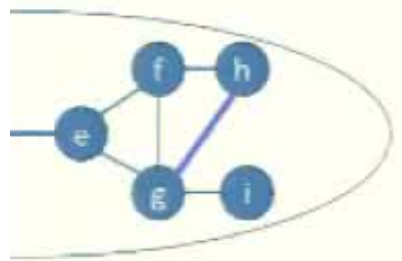
Score({a,b})
did **not** change

$$\text{score}(\mathcal{M}, \{a, b\}) = \text{score}'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$



The addition of edge {a,b} need **not** be re-examined in the new model

1}

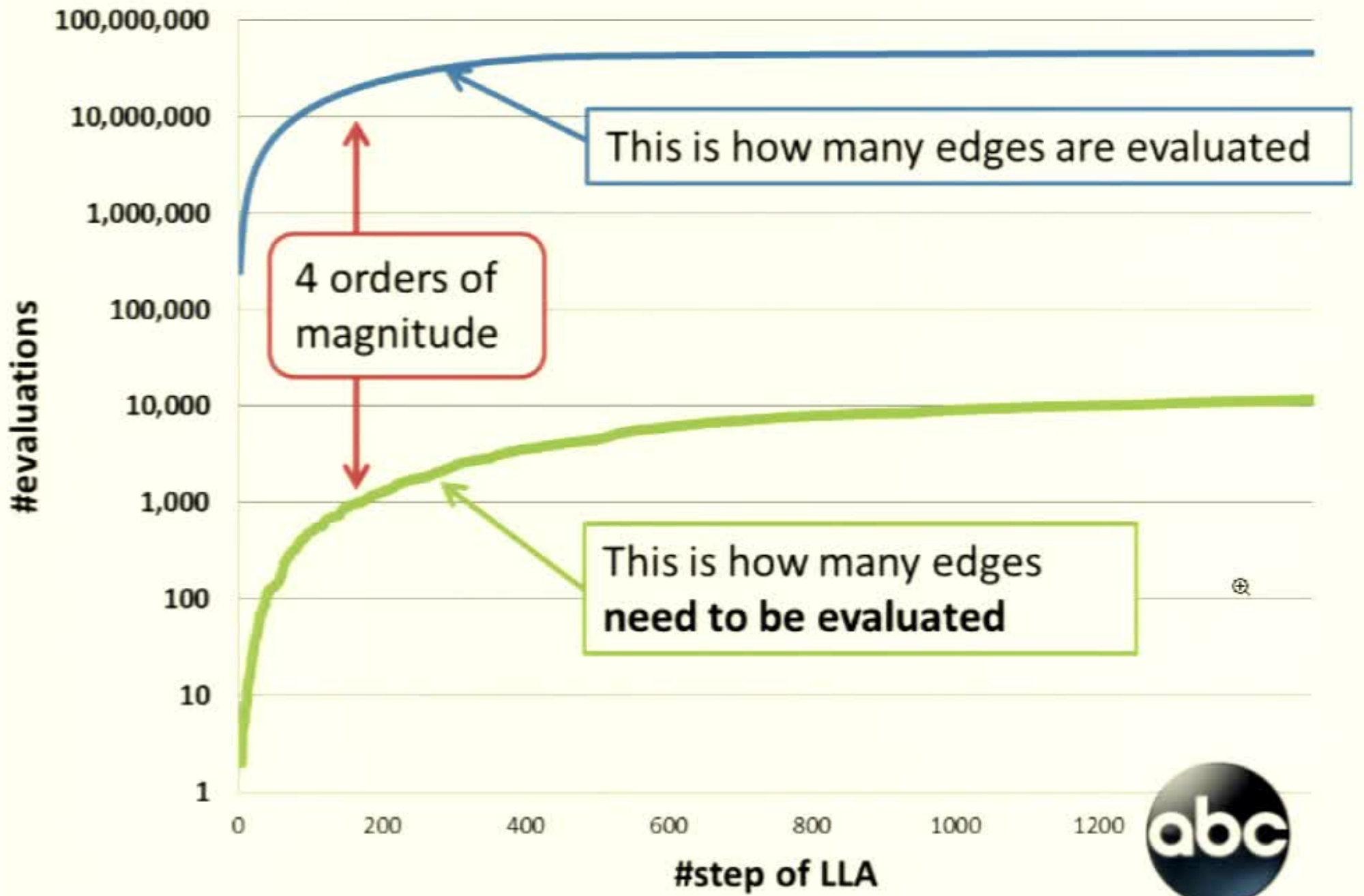


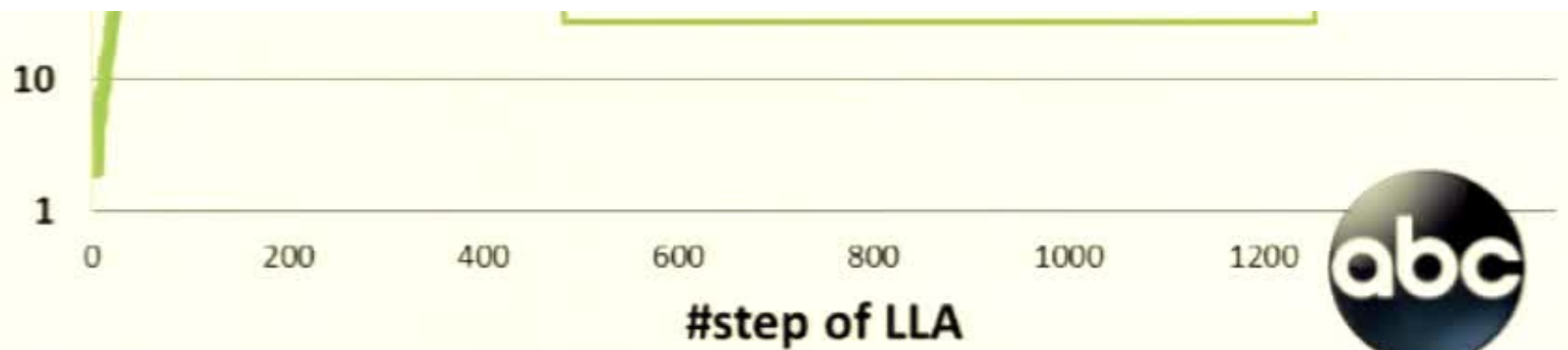
Score({a,b})
did **not** change

$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\}, \{a, c, d\}, \{b, c, d\}, \{c, d\})$$

Deletion of edge {a,b} need **not** be re-
quired in the new model







We know: if S_{ab} does not change between different modifications of the graph, then the addition of $\{a,b\}$ need not be re-examined

1. Use a data structure that gives direct

We know: if S_{ab} does not change between different modifications of the graph, then the addition of $\{a,b\}$ need not be re-examined



1. Use a data structure that gives direct access to minimal separators for every potential edge



2. Keep track of the minimal separators for every potential edge



3. Maintain an ordered list of all the potential edges (priority queue)





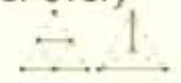
ion?



We know: if S_{ab} does not change between different modifications of the graph, then the addition of $\{a,b\}$ need not be re-examined



1. Use a data structure that gives direct access to minimal separators for every potential edge



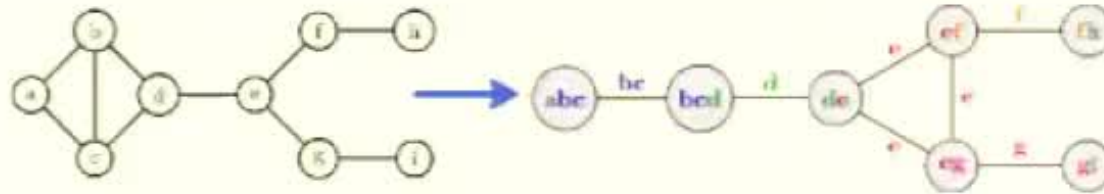
2. Keep track of the minimal separators for every potential edge



3. Maintain an ordered list of all the potential edges (priority queue)

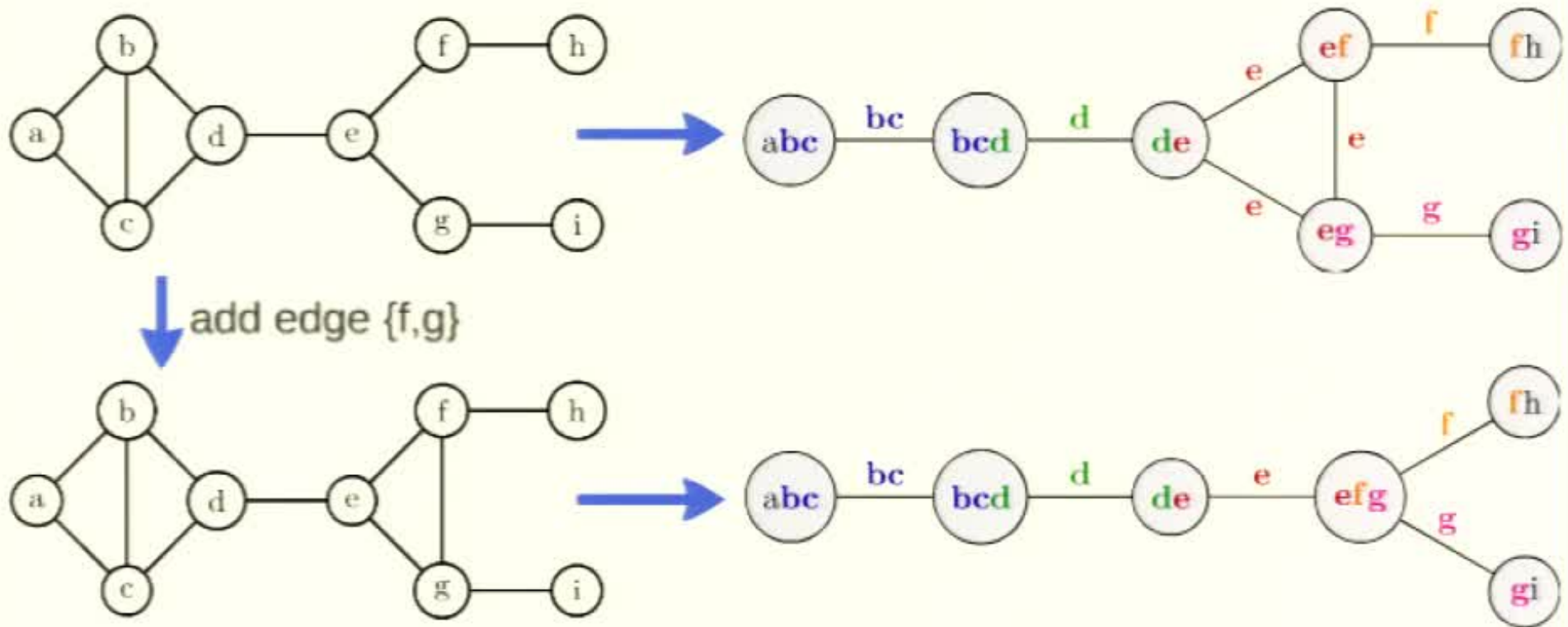


Clique graph



[1]: A. Deshpande et al., "Efficient stepwise selection in decomposable models," in *UAI 2001*.

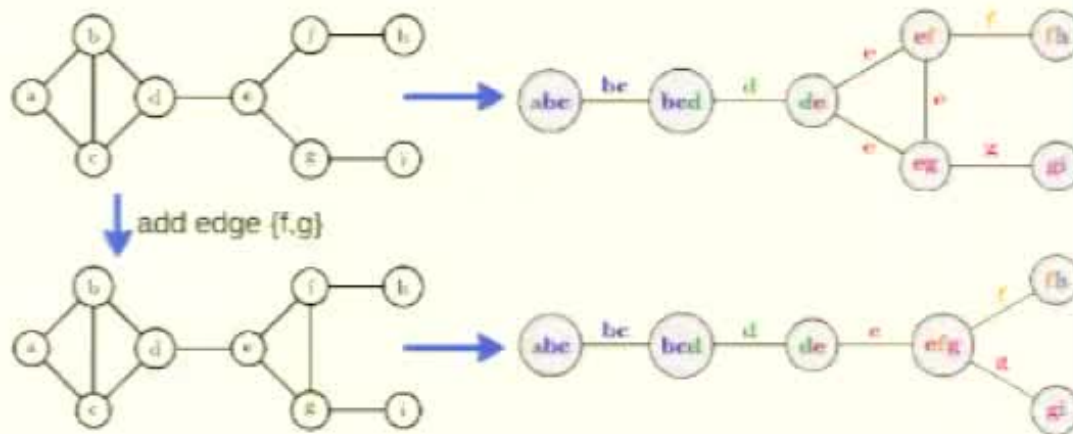
Clique graph



There are algorithms that can directly update the clique-graph [1]

BUT those algorithms cannot track the minimal separators S_{ab}
→ We make this possible - *details in the paper*

Clique graph

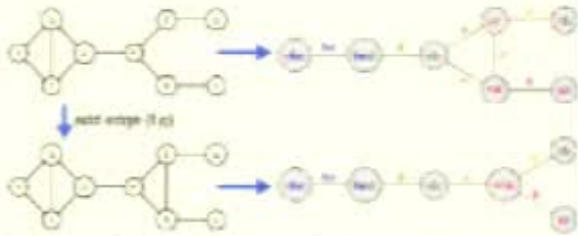


There are algorithms that can directly update the clique-graph [1]

BUT those algorithms cannot track the minimal separators S_{ab}
→ We make this possible - *details in the paper*

[1]: A. Deshpande et al., "Efficient stepwise selection in decomposable models," in UAI 2001.

Clique graph



There are algorithms that can directly access the clique graph [1]

BUT these algorithms cannot track the minimal separators $S_{i,j}$

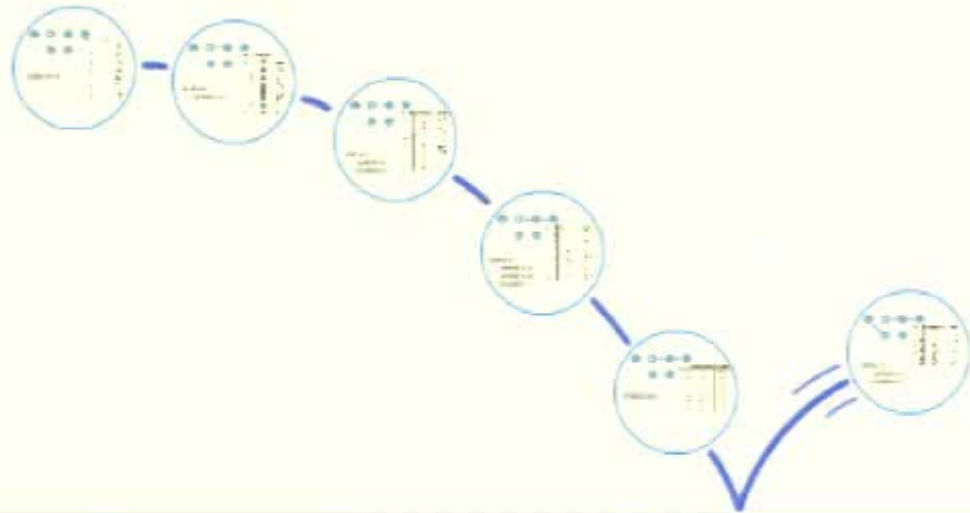
→ We make this possible - details in the paper

[1] A. Desrosiers et al. "Efficient algorithms for inference in decomposable models." in IJAI 2002

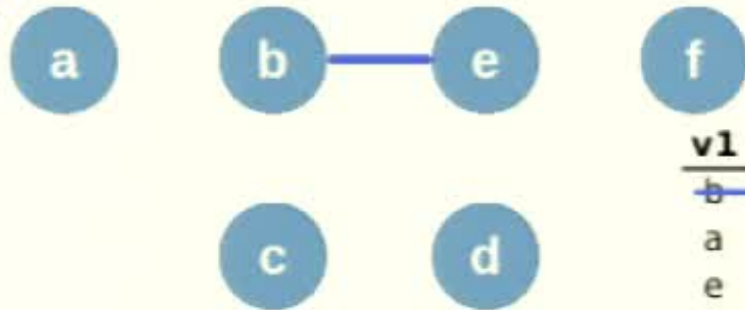
v^2



$v \cdot \log(v)$

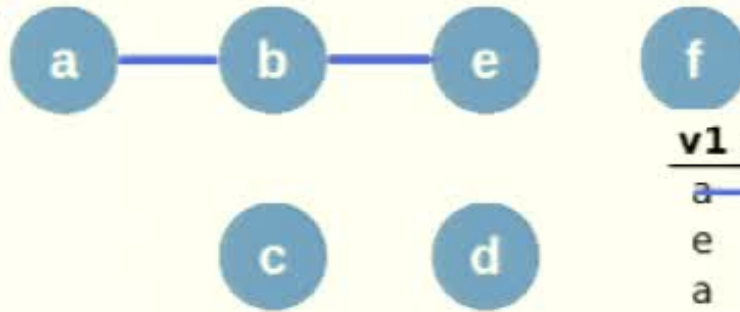


Priority queue



Add b-e

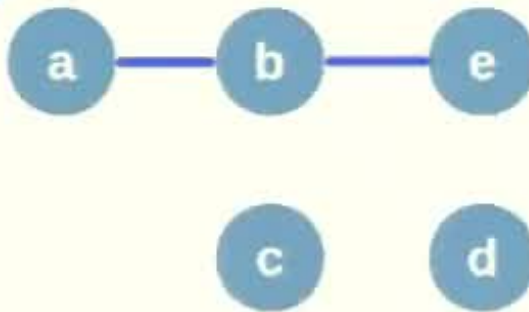
v1	v2	separator	score
b	e	{}	96.2
a	b	{}	72.8
e	f	{}	60.9
a	e	{}	49.5
a	f	{}	42.8
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6



v1	v2	separator	score
a	b	{}	72.8
e	f	{}	60.9
a	e	{}	49.5 ④
a	f	{}	42.8
b	e	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add a-b
 • update a-e

b



v1	v2	separator	score
e	f	{}	60.9
a	f	{}	42.8
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
a	e	{b}	12.4
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add e-f

- update b-f
- disable a-f

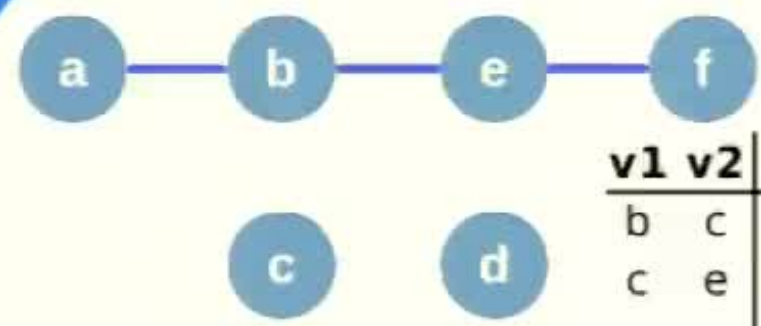


v1	v2	separator	score
e	f	{}	60.9
a	f	{}	42.8
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
a	e	{b}	12.4
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add e-f

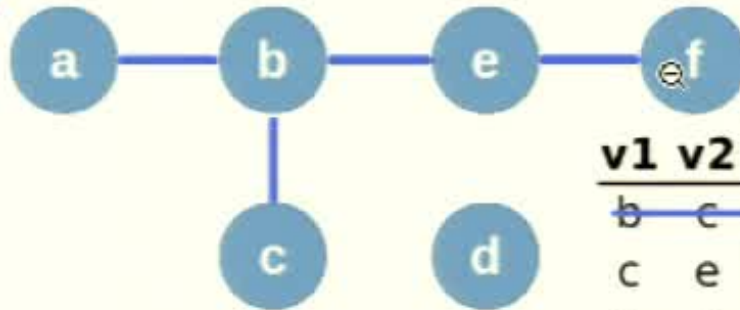
- update b-f
- disable a-f

e	{b}	12.4
f	{}	8.1
e	{}	7.3
f	{}	4.8
f	{}	4.6



v1	v2	separator
b	c	{}
c	e	{}
a	c	{}
c	d	{}
b	f	{e}
b	c	{}
a	e	{b}
c	f	{}
d	e	{}
d	f	{}

- Add b-c
- update a-c
 - update c-e

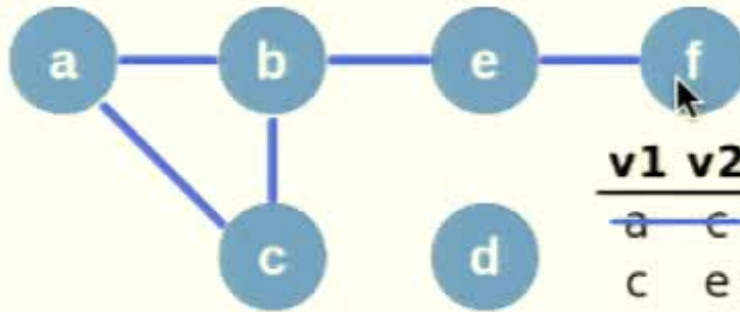


v1	v2	separator	score
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add b-c

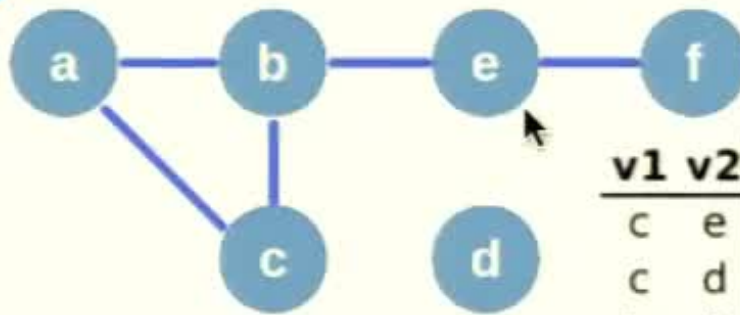
- update a-c
- update c-e
- disable c-f

b



Add a-c

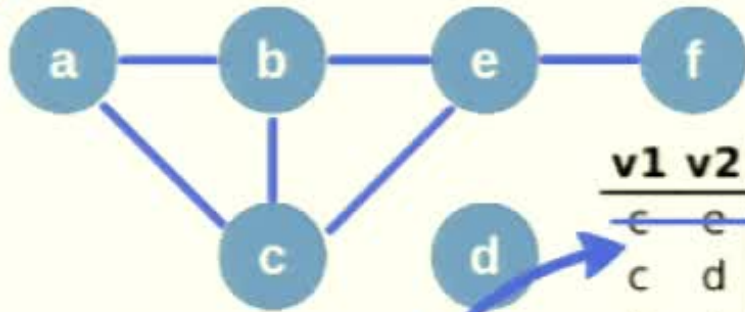
v1	v2	separator	score
a	c	{b}	84.5
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6



v1	v2	separator	score
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add c-e

- update a-e
- enable c-f



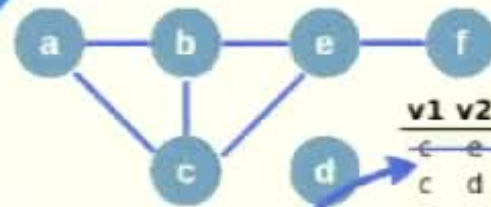
c f | {e} | 49.4

v1	v2	separator	score
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

C

Add c-e

- update a-e
- enable c-f



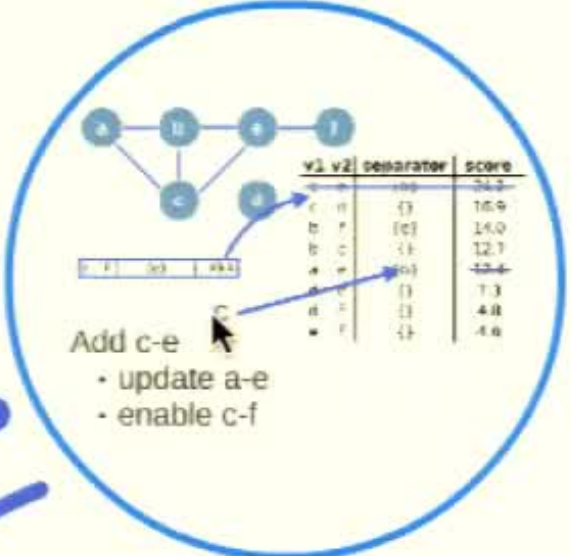
C ⊕

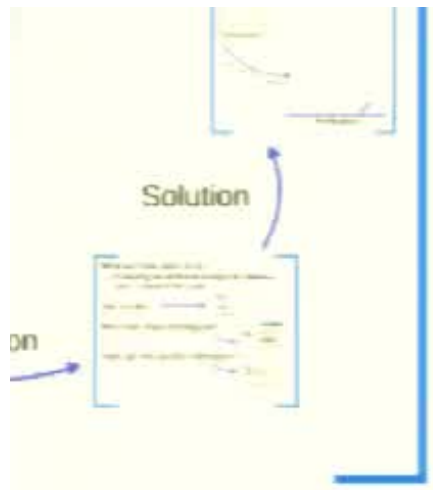
v1	v2	separator	score
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add c-e

- update a-e
- enable c-f

separator	score
(a)	24.2
()	16.0
(e)	14.0
()	12.7
(b)	12.4
()	7.3
()	4.8
()	4.6



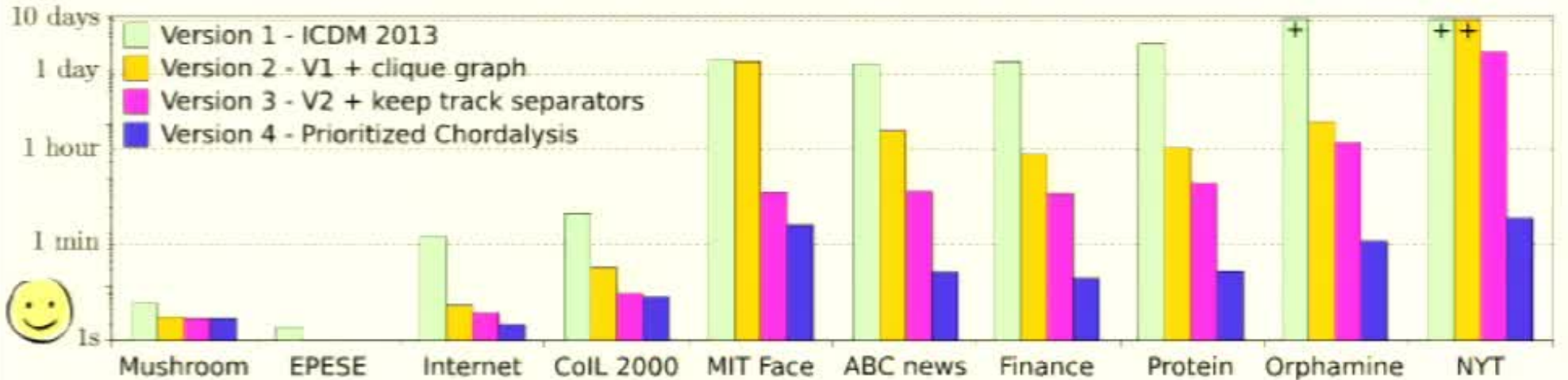


It works!

itized
analysis



Running times



#Vars 20 25 70 85 300 500 500 700 1,200 2,000



Take-home message



Prioritized Chordalysis:

1. can analyse data with 1,000+ variables
2. does not sacrifice the soundness
3. is released on **GitHub**



<https://github.com/fpetitjean/>

Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb



Thanks for your attention!



<http://www.francois-petitjean.com>



francois.petitjean@monash.edu



@LeDataMiner



Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb



2015 SIAM International
Conference on **DATA MINING**

