Causal Effect Evaluation and Causal Network Learning

Zhi Geng

Peking University, China

June 25, 2014

Outline

1. Causal Effect Evaluation
   - Yule-Simpson paradox
   - Causal effects
   - Surrogate and surrogate paradox
2. Causal Network Learning
   - Decomposing learning
   - Active learning
   - Local learning

Outline

Outline

Yule-Simpson paradox

"Human can be compared to
a frog at the bottom of a well"

Frog's sight ⇒

Frog ⇒

Can the frog make
a correct inference
about the universe
from its sight?

Yule-Simpson Paradox (Yule, 1900; Simpson, 1951)

	Cancer	Control	Total
Smoking	100	100	200
Non-smoking	80	120	200
	$RD = \frac{100}{200} - \frac{80}{200} = 0.10$		

	Male (Gene=+)		Female (Gene=−)	
	Cancer	Control	Cancer	Control
Smoking	90	60	10	40
Non-smok	35	15	45	105
	$RD_M = \frac{90}{150} - \frac{35}{50} = -0.10$		$RD_F = \frac{10}{50} - \frac{45}{150} = -0.10$	

Smoking is bad for humans, but good for both men and women, called Yule-Simpson paradox.

It is because we used an association measurement.

Outline

Definitions of Causal Effects (Neyman, 1923; Rubin, 1974)

- For an individual i,
  $Y_1(i)$: potential outcome if treatment T were 1 (Smoking),
  $Y_0(i)$: potential if treatment T were 0 (Non-smoking),
- Observed outcome:
$$Y(i) = \begin{cases} Y_1(i), & T(i) = 1; \\ Y_0(i), & T(i) = 0. \end{cases}$$
- Individual Causal Effect:
$$ICE(i) = Y_1(i) - Y_0(i).$$
  Only one of $Y_1(i)$ and $Y_0(i)$ is observable for a person i.
- Average Causal Effect (ACE):
$$ACE(T \to Y) = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

Causal effect $\neq$ Association measure

- Generally, ACE is not identifiable.

$$ACE(T \to Y) \neq RD.$$

- But for a randomized study, we have $(Y_1, Y_0) \perp\!\!\!\perp T$. Thus

$$\begin{aligned} ACE(T \to Y) &= E(Y_1) - E(Y_0) \\ &= E(Y_1|T=1) - E(Y_0|T=0) \\ &= E(Y|T=1) - E(Y|T=0) \\ &= RD, (\text{ An association measure}). \end{aligned}$$

We can evaluate ACE using association measures even if there are unobserved variables like a frog in a well.

Observational Studies

- For an observational study, we require the ignorable treatment assignment assumption $(Y_1, Y_0) \perp\!\!\!\perp T | X$,
  where X is a sufficient confounder set.
  If X is observed, then

$$ACE(T \to Y) = \sum_x ACE(T \to Y|x)P(x).$$

  No Yule-Simpson paradox for ACE:

$$ACE(T \to Y|x) > 0, \forall x \Longrightarrow ACE(T \to Y) > 0.$$

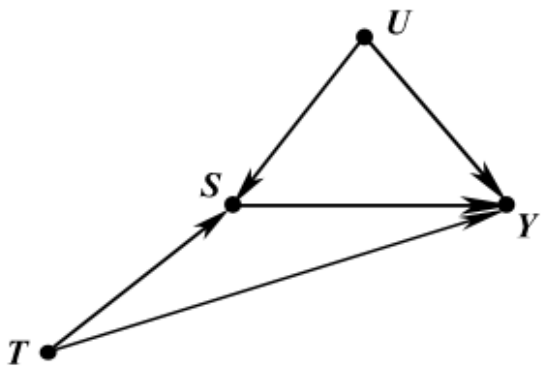
  Many approaches are used for estimating ACE:
  Stratification, Propensity score, Inverse probability weighting, ...
- If X is unobserved, we need to find
  an instrumental variable (IV) Z ($Z \not\perp\!\!\!\perp T$ and $Z \perp\!\!\!\perp X$),
  to estimate ACE.

Outline

Surrogate: a scapegoat (替罪羔羊)

- When it is difficult to observe the endpoint variable, instead, we often observe a surrogate variable (or biomarker).
- For example, it may take too long time to observe the survival times (e.g., 5 years) for AIDS patients. Thus CD4 count is often used as a surrogate for the survival time in a clinical trial of AIDS treatment.

Criteria for selecting surrogates



Notation:

- T: Treatment (randomized),
- Y: The endpoint variable,
- S: Surrogate (an intermediate variable),
- U: Unobserved confounder (S not randomized),
- S_t: potential outcome of S if treatment were t.
- Y_{st}: potential outcome of Y if $T = t$ and $S = s$.

Criteria for surrogates

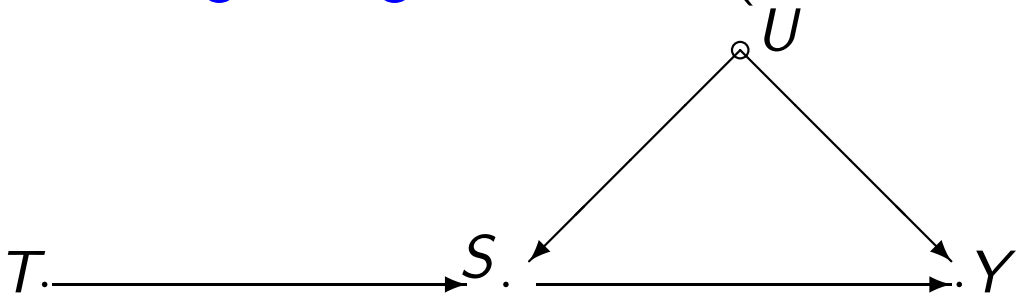
There have been many criteria for selecting a surrogate:

1. A strong correlation surrogate criterion:
   A surrogate should strongly correlate to the endpoint.
2. The conditional independence criterion (Prentice, 1989):
   A surrogate should break all association between T and Y, $Y \perp\!\!\!\perp T | S$.
3. The principal surrogate criterion (Frangakis & Rubin, 2002):
   A surrogate should satisfy the property of causal necessity:
   No effect on surrogate $\Rightarrow$ No effect on endpoint

$$S_{T=1}(u) = S_{T=0}(u) \implies p(Y_{T=0}) = p(Y_{T=1}), \text{ for these } u.$$

Criteria for Surrogates

- The strong surrogate criterion (Lauritzen, 2004):



  where U is an unobserved variable.

- A surrogate S should break the causal path from T to Y.
  No causal effect of T on $S \Longrightarrow$ no causal effect of T on Y.
  Thus a strong surrogate is also a principal surrogate.

Surrogate paradox

- We pointed out that for all of the above criteria for surrogates, it is possible that
  treatment T has a positive effect on surrogate S,
  which in turn has a positive effect on endpoint Y,
  but T has a negative effect on endpoint Y.

$$T \xrightarrow{ACE(T \to S) = +} S \longrightarrow Y$$

Surrogate paradox

- We pointed out that for all of the above criteria for surrogates, it is possible that
  treatment T has a positive effect on surrogate S,
  which in turn has a positive effect on endpoint Y,
  but T has a negative effect on endpoint Y.

$$T \xrightarrow{ACE(T \to S) = +} S \xrightarrow{ACE(S \to Y) = +} Y$$

Surrogate paradox

- We pointed out that for all of the above criteria for surrogates, it is possible that
  treatment T has a positive effect on surrogate S,
  which in turn has a positive effect on endpoint Y,
  but T has a negative effect on endpoint Y.

$$\begin{tikzcd} T \arrow[r, "ACE(T \to S) = +"] \arrow[rr, bend right, "ACE(T \to Y) = -"'] & S \arrow[r, "ACE(S \to Y) = +"] & Y \end{tikzcd}$$

Surrogate paradox

- We pointed out that for all of the above criteria for surrogates, it is possible that
  treatment T has a positive effect on surrogate S,
  which in turn has a positive effect on endpoint Y,
  but T has a negative effect on endpoint Y.

$$\begin{tikzcd} T \arrow[r, "ACE(T \to S) = +"] \arrow[rr, bend right, "ACE(T \to Y) = -"'] & S \arrow[r, "ACE(S \to Y) = +"] & Y \end{tikzcd}$$

- We call this a surrogate paradox (Chen, G & Jia, 2007).

A real example

Moore (2005)'s book: *"Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster"*

- Doctors have the knowledge on irregular heartbeats:
  - irregular heartbeat is a risk factor for sudden death,
  - correcting irregular heartbeats would prevent sudden death.
- Thus 'correction of heartbeat' as a surrogate, several drugs (Enkaid, Tambocor, Ethmozine) were approved by FDA.
- But a later CAST study showed: the correction of heartbeat did not improve survival times but increased mortality.

Numerical example

- T: treatment ($T = 1$ treated, $T = 0$ control),
- S: Correction of irregular heartbeat ($S = 1$ corrected, $S = 0$ not),
- Y: the survival time.

Assume

1. all effects of treatment T on survival Y are through intermediator S, that is, $Y_{st} = Y_{st'} = Y_s$,
2. correction of heartbeat can increase survival time for every patient u

$$Y_{s=0}(u) < Y_{s=1}(u).$$

Numerical example (continued)

Group	No.	$S_{T=0}$	$S_{T=1}$	$Y_{S=0}$ <	$Y_{S=1}$	$Y_{T=0}$	$Y_{T=1}$
1	20	0	0	9	10	9	9
2	40	0	1	6	7	6	7
3	20	1	0	5	8	8	5
4	20	1	1	3	5	5	5

$$ACE(T \to S) = \frac{40+20}{100} - \frac{20+20}{100} = \frac{20}{100} > 0,$$

but

$$ACE(T \to Y) = \frac{9 \times 20 + 7 \times 40 \cdots}{100} - \frac{\cdots + 5 \times 20}{100} = 6.6 - 6.8 < 0.$$

Correction of heartbeats S is not a valid surrogate.

Criteria for Surrogates

- Generally for a continuous or ordinal Y, define the distributional causal effect (DCE) by

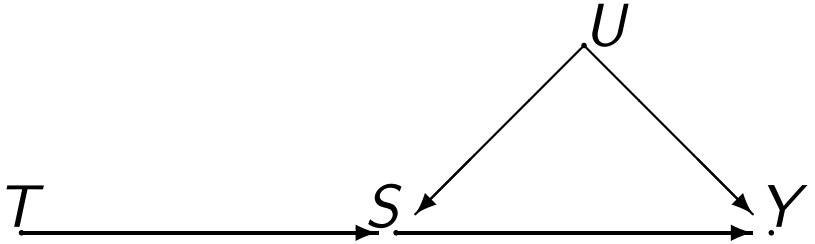
$$DCE[T \to (Y > y)] = P(Y_{T=1} > y) - P(Y_{T=0} > y).$$

$$DCE[T \to (S > s)] = P(S_{T=1} > s) - P(S_{T=0} > s).$$

- Goal: Without observing Y, but observing S instead, we want to predict the sign $(+, -, 0)$ of $DCE[T \to (Y > y)]$ using the sign of $DCE[T \to (S > s)]$.
- To avoid the surrogate paradox, we give different conditions, some are based on associations, and some are based on causations.

Causation-based Criteria for Surrogate

- **Theorem 1.** (Ju and G, JRSS B, 2010)
  Assume that the causal network is true: without $T \longrightarrow Y$



  If

  1. the DCEs of S on Y conditional on $U = u$ have the same sign for all u, and
  2. the DCEs of T on S conditional on $U = u$ have the same sign for all u.

  then the sign of $DCE[T \to (Y > y)]$ can be predicted by the sign of $DCE[T \to (S > s)]$.

- These conditions cannot be tested by data even Y is observed because U is unobserved.

Association-based criteria

- We propose association-based conditions.

- **Theorem 2.** (Wu, He and G, 2011, Statist Med)
  If
  1. $P(Y > y|s, T = 1)$ or $P(Y > y|s, T = 0)$ monotonically increases in s and
  2. $P(Y > y|s, T = 1) \geq P(Y > y|s, T = 0)$ for all s,

  then

$$DCE[T \to (S > s)] \geq 0 \Longrightarrow DCE[T \to (Y > y)] \geq 0$$

- The conditions are testable if Y is observed in a validation study.
- But the reverse '$\Longleftarrow$' is not true.

Equivalence relationships of CE's signs

- **Theorem 3.** If
  1. Prentice's criterion $Y \perp\!\!\!\perp T | S$,
  2. $P(Y > y|s)$ increases in s and
  3. S is from an exponential family conditional on T,

  then

$$\begin{aligned} & Sign[ACE(T \to S)] = Sign[DCE(T \to S)] \\ = \quad & Sign[ACE(T \to Y)] = Sign[DCE(T \to Y)], \end{aligned}$$

  where *Sign* means '$= 0$', '> 0' or '< 0'.

Summary of criteria for surrogates

- The principal surrogate and the strong surrogate: only

$$CE(T \to S) = 0 \Longrightarrow CE(T \to Y) = 0.$$

- The monotonicity: further

$$CE(T \to S) \geq (\leq) 0 \Longrightarrow CE(T \to Y) \geq (\leq) 0.$$

- Prentice's criterion and S from the exponential family : equivalence relationships

$$CE(T \to S) > (<, =)\ 0 \Longleftrightarrow CE(T \to Y) > (<, =)\ 0.$$

Outline

Causal network, DAG

- Causal relationships among variables can be represented by a directed acyclic graph (DAG) (Pearl, 2000):

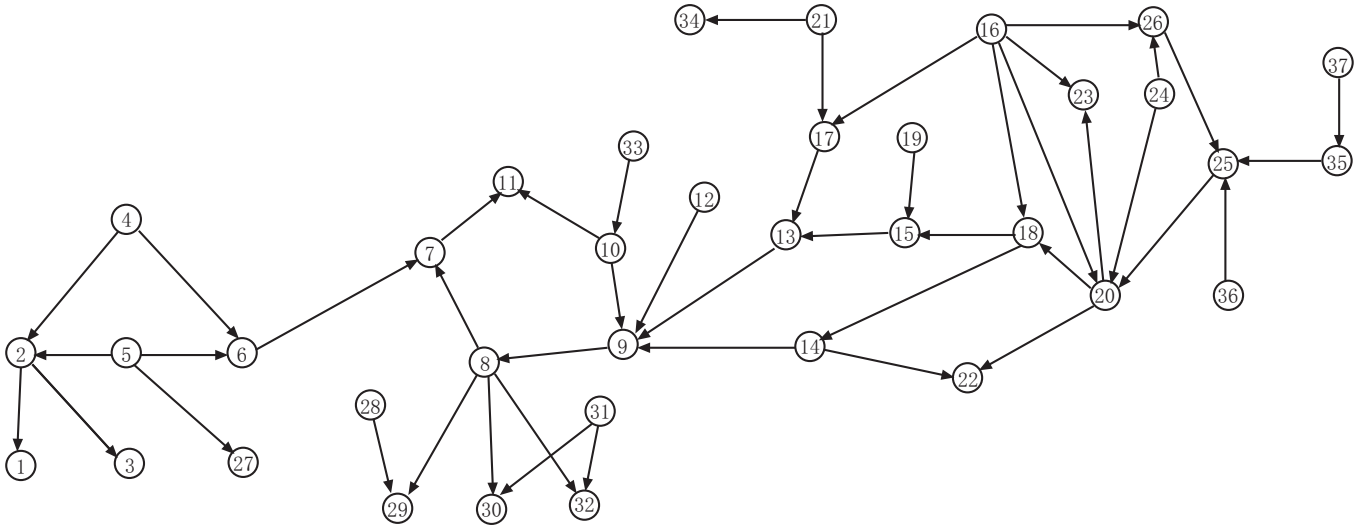


Figure: ALARM: a medical diagnostic network (Belinlich et al., 1989)

Three proposed approaches

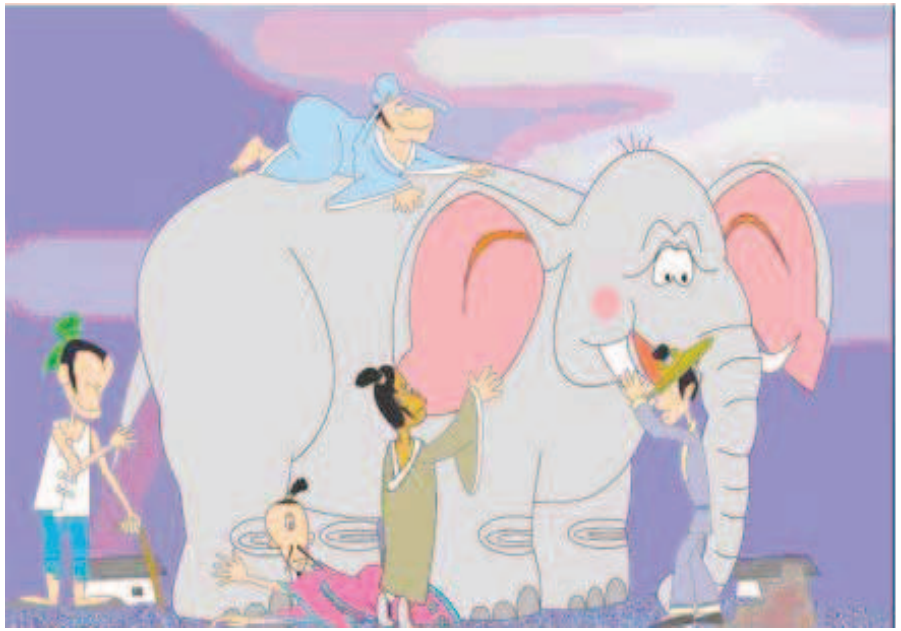
We propose three approaches for learning networks from data:

- Decomposing learning:
  - Learn local networks from incomplete data and combine them,
  - Recursively decompose a large network learning to several smaller networks learning;
- Active learning:
  Manipulate some variables to change an association network to a causation network;
- Local learning:
  Learn a local structure around a target variable of interest.

Outline

Blind men touch an elephant (盲人摸象)

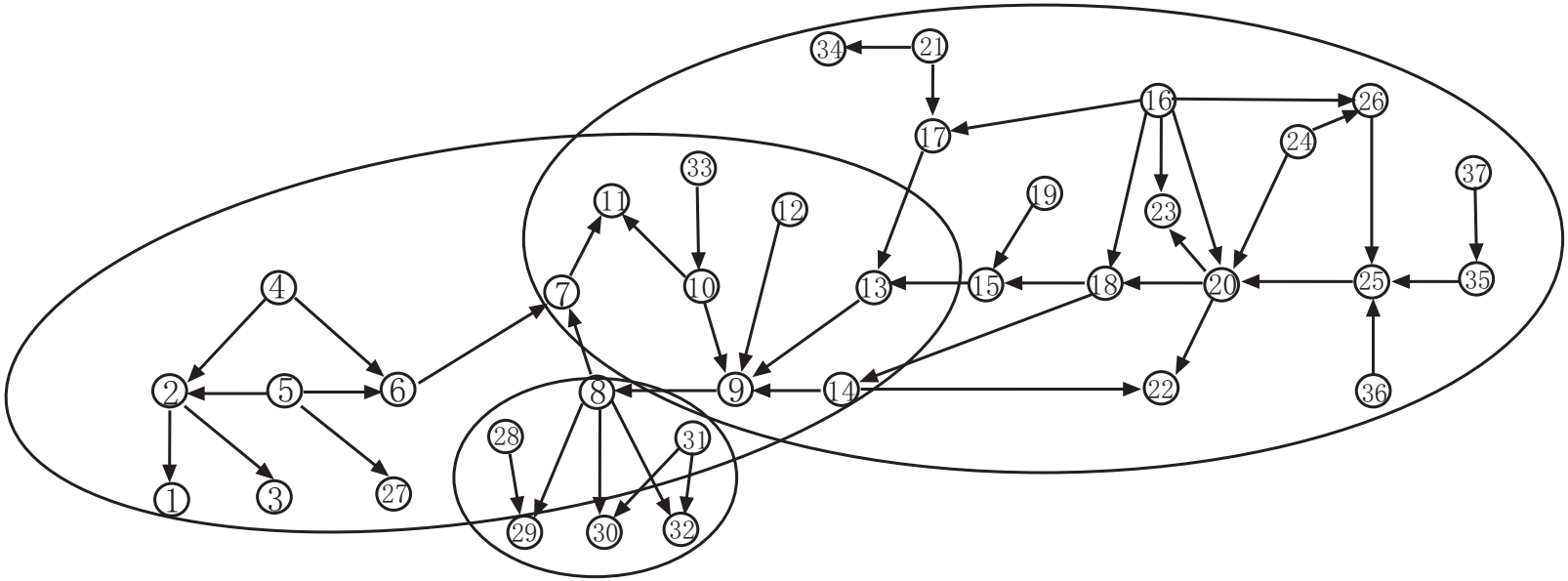
We discuss how blind men can discover an elephant:



(Xie, G and Zhao, 2006, Artificial Intelligence)

Decomposing learning

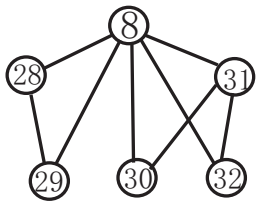
The decomposing approach:



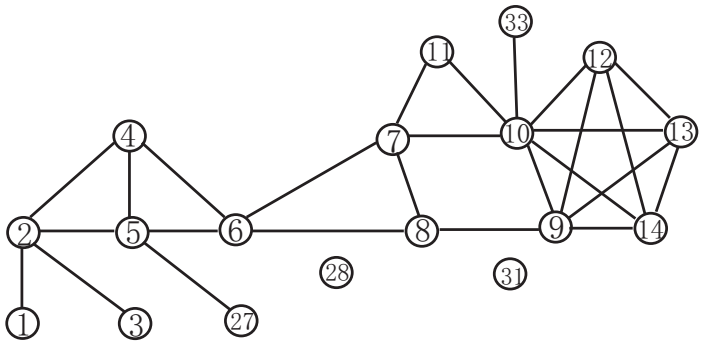
- Three experts in different areas observed different variable sets.
- We obtained 3 incomplete data sets of the variable sets.

Decomposing learning

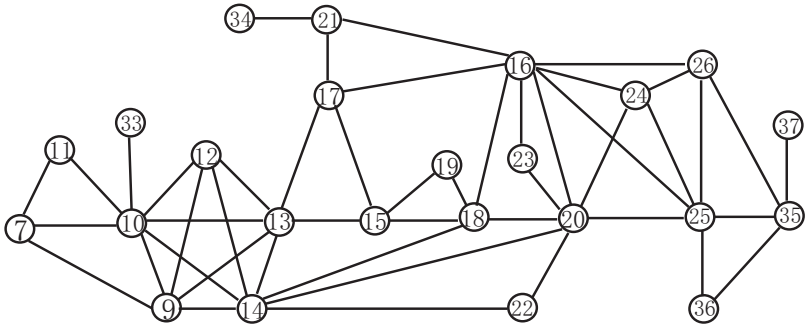
Learn undirected subgraphs from each data set:



(a) from data set 1



(b) from data set 2

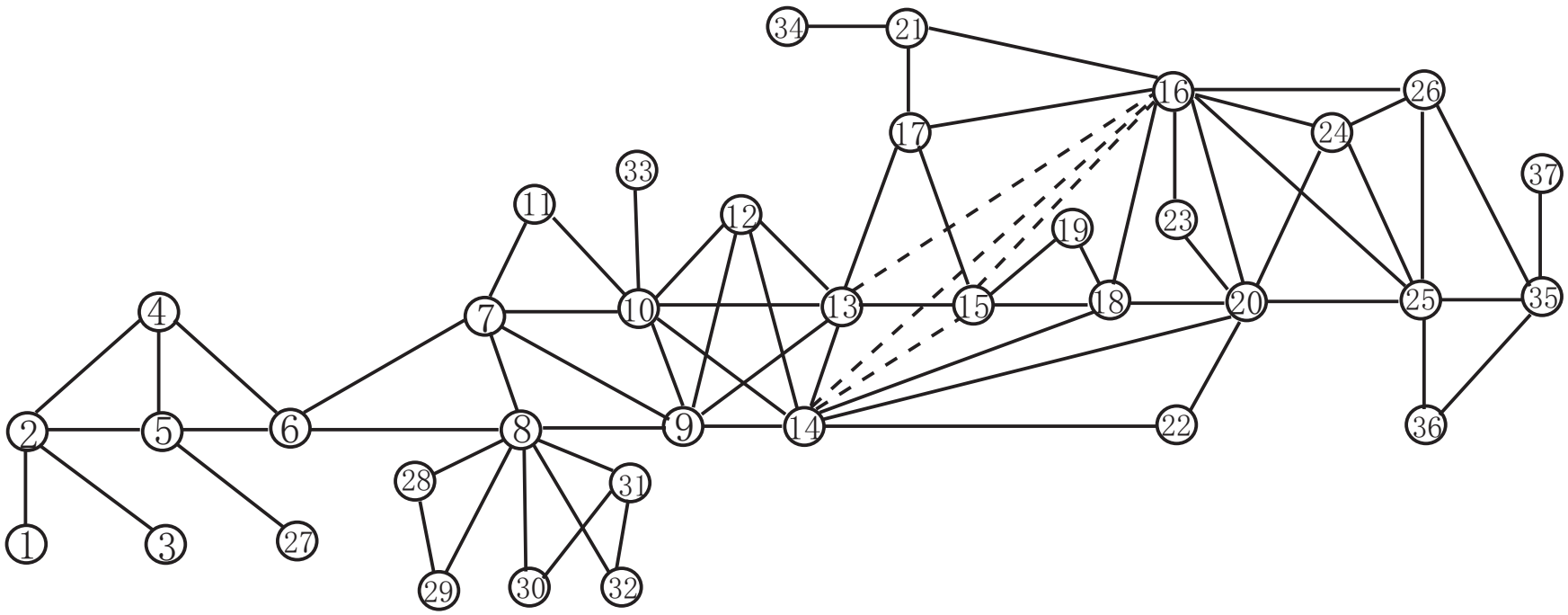


(c) from data set 3

Some edges $(7-9)$ may be spurious due to incomplete data.

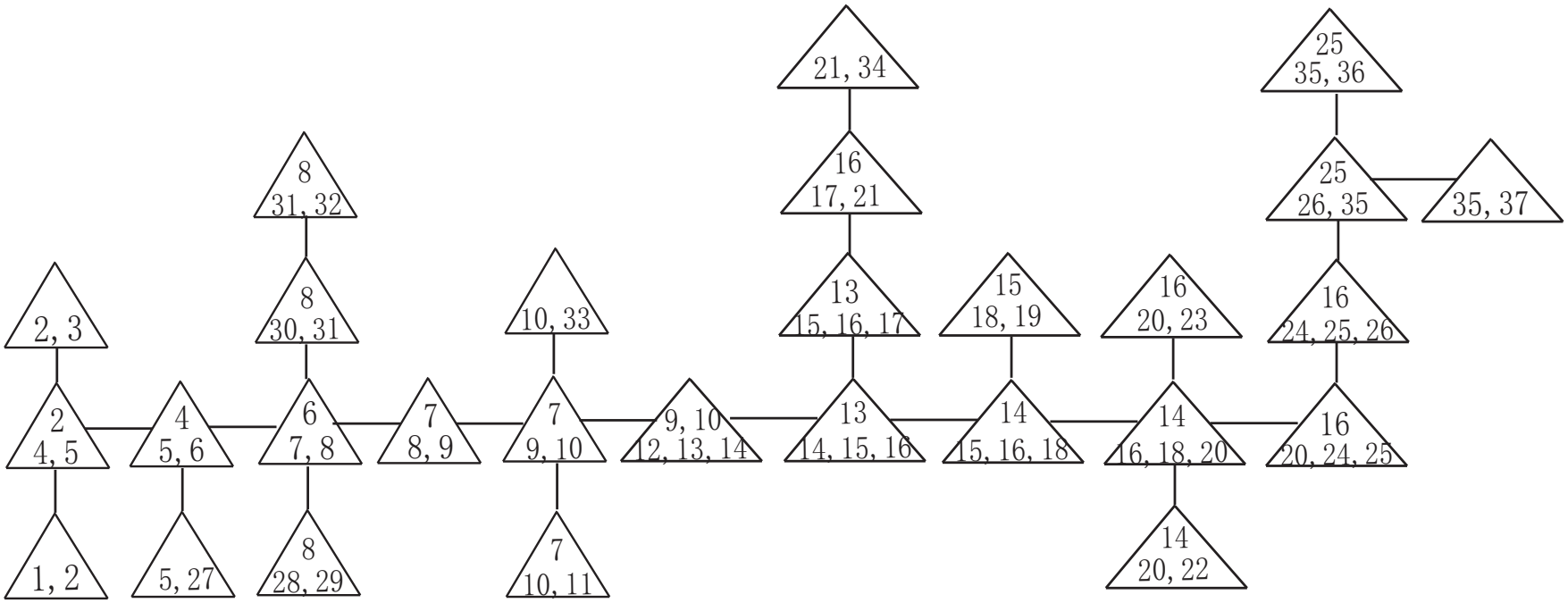
Decomposing learning

Combine these subgraphs together,
triangulate it by adding dashed edges:



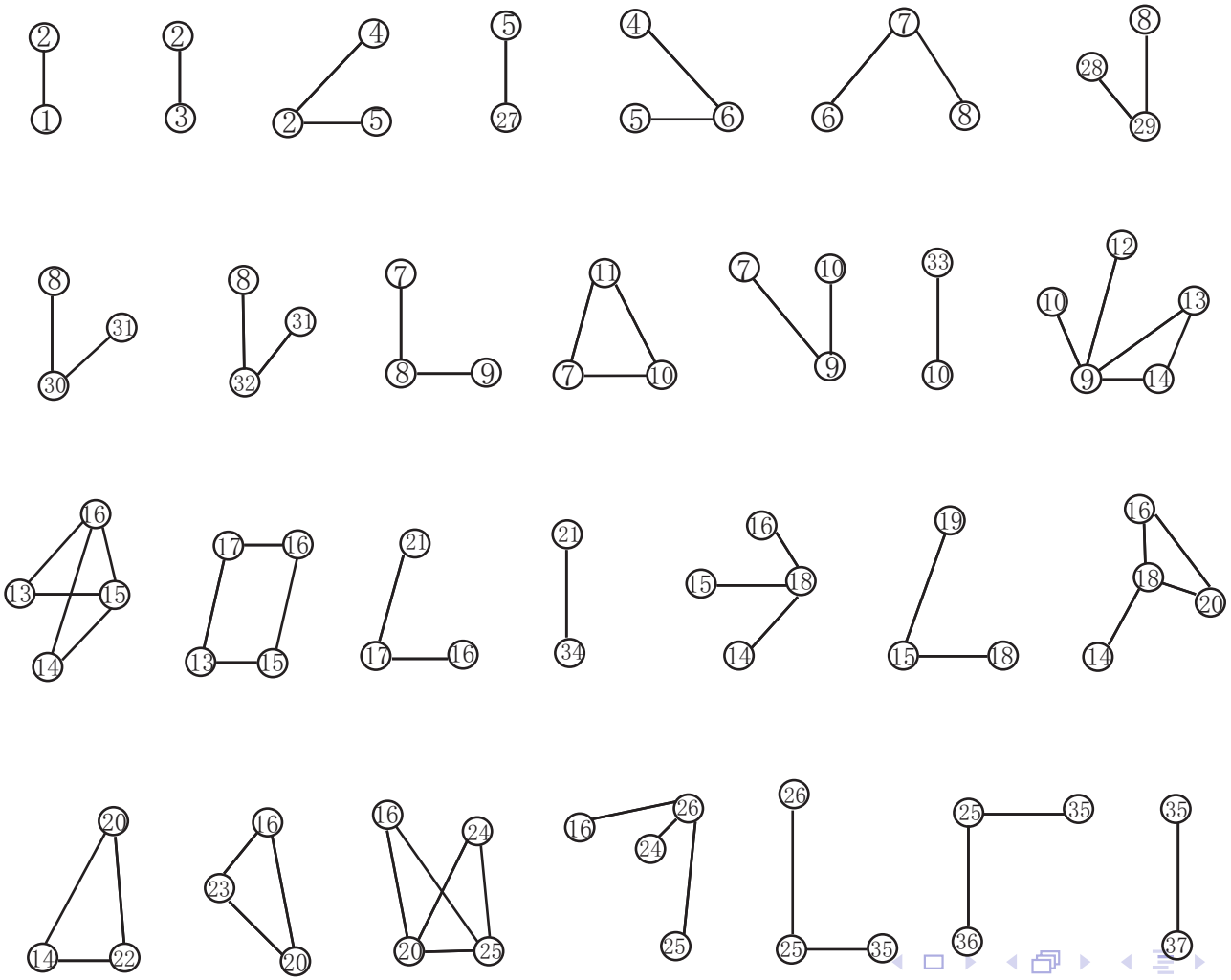
Decomposing learning

Construct the separation tree,
each (node) cluster represents a complete subgraph,
the largest cluster has only 5 variables:



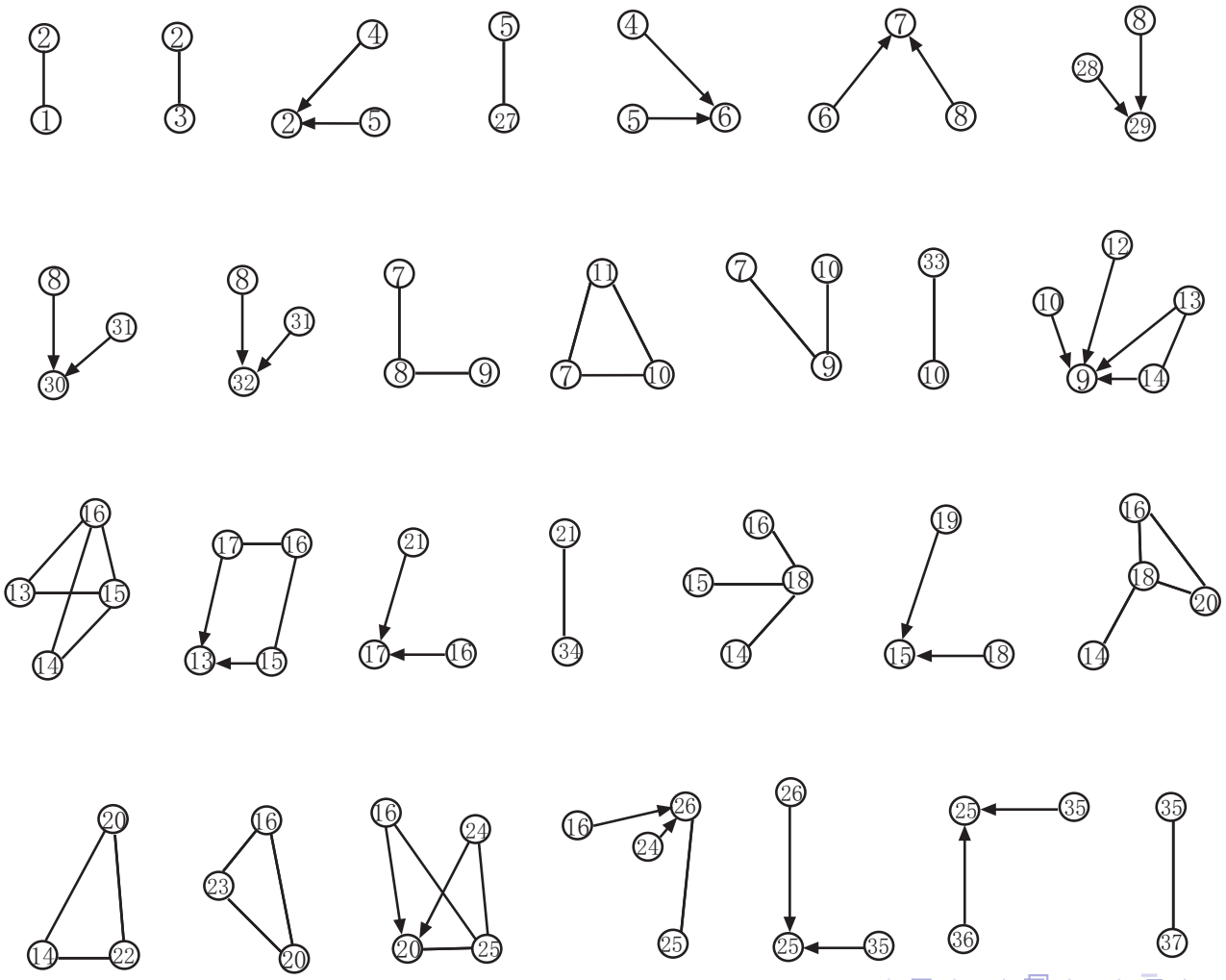
Decomposing learning

Re-construct undirected subgraphs in each cluster:



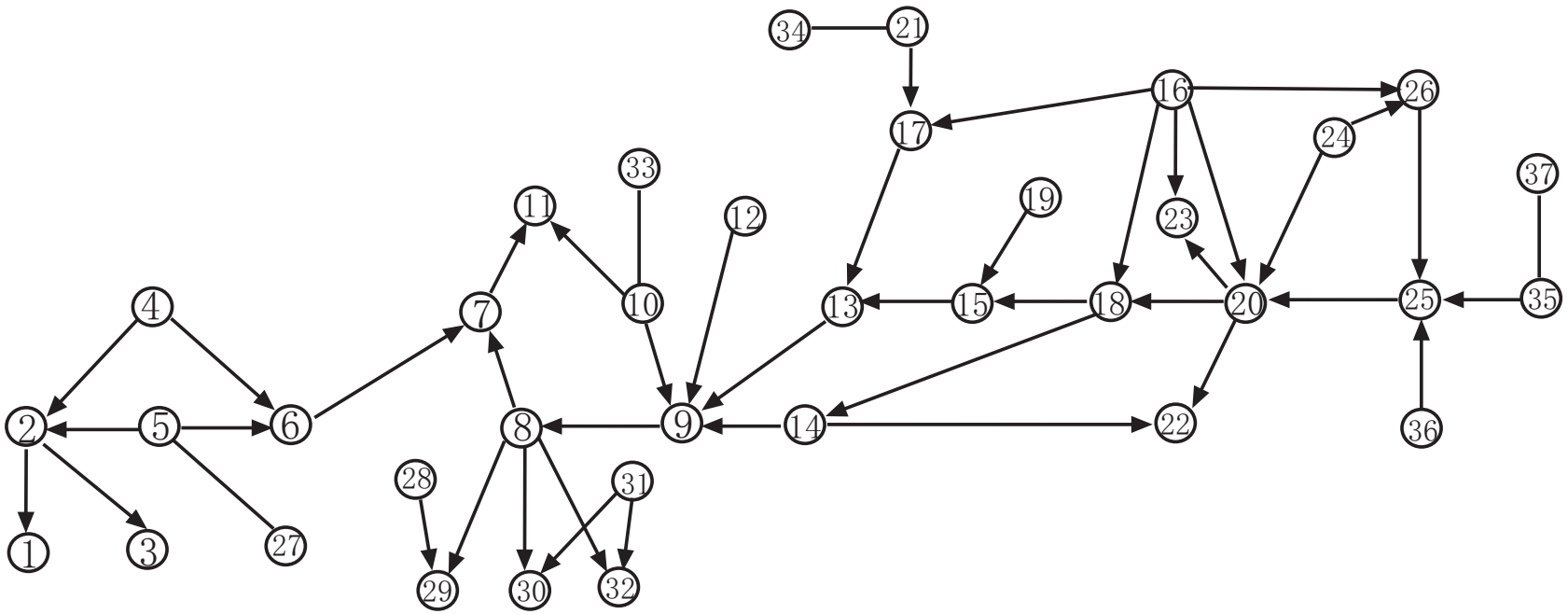
Decomposing learning

Orient edges in each subgraph:



Decomposing learning

Combining subgraphs and orienting other undirected edges, we obtain the Markov equivalence class:



Recursive learning

- A recursive learning approach by divide and conquer. (Xie and G, 2008, JMLR)
- It recursively decomposes a problem of learning a large graph into problems of learning two small graphs.

Recursive Learning

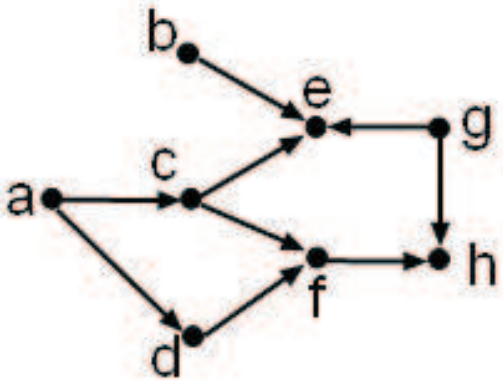
PROCEDURE DecompLearning (K, $\bar{L}_K$)

1. Construct an undirected independence graph $\bar{G}_K$;
2. If $\bar{G}_K$ has a decomposition (A, B, C) (i.e., $A \perp\!\!\!\perp B | C$)
   Then
   - DecompLearning ($A \cup C$, $\bar{L}_{A\cup C}$);
   - DecompLearning ($B \cup C$, $\bar{L}_{B\cup C}$);
   - Set $\bar{L}_K$ = CombineSubgraphs ($\bar{L}_{A\cup C}$, $\bar{L}_{B\cup C}$)

   Else
   - Construct the local skeleton $\bar{L}_K$ directly from data (e.g. the IC algorithm).
3. RETURN ($\bar{L}_K$).

Example

Data are generated from the unknown causal network:



Top-down stage

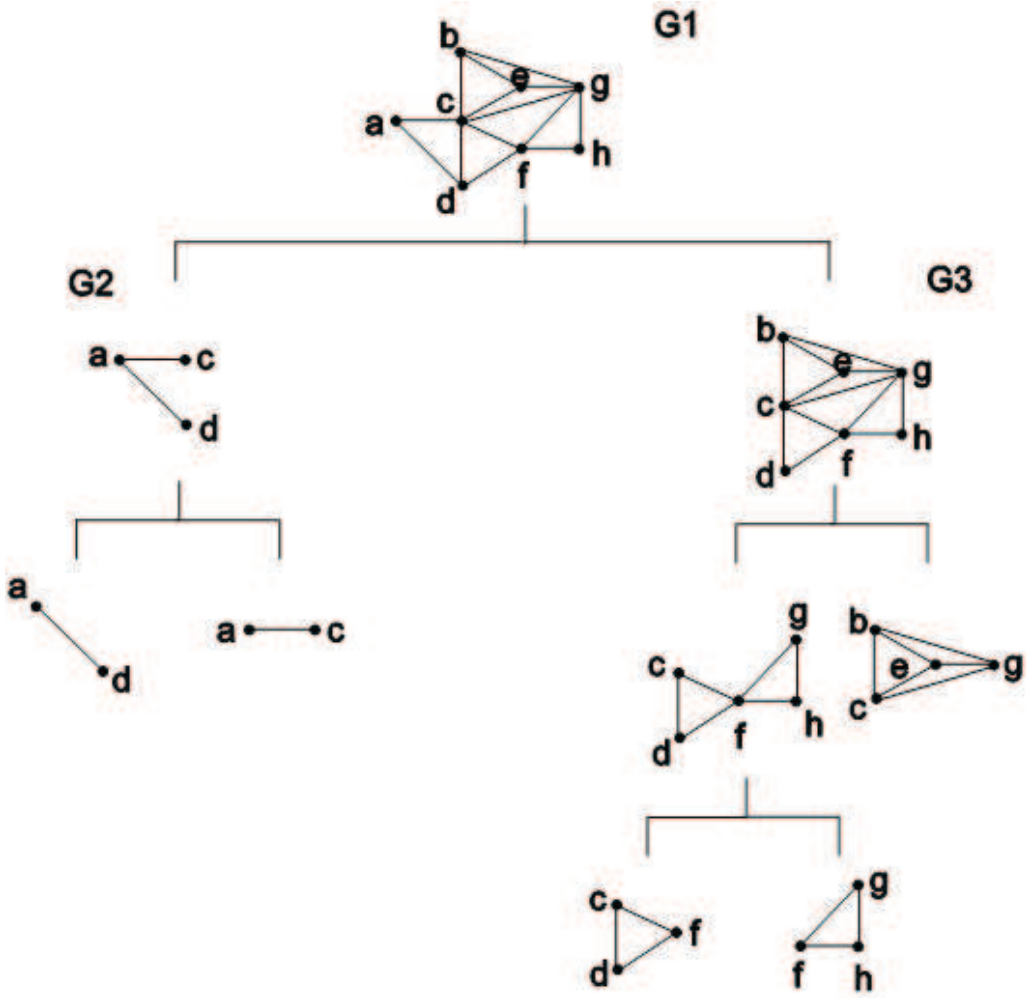


Figure: The tree obtained at the top-down step.

Top-down stage

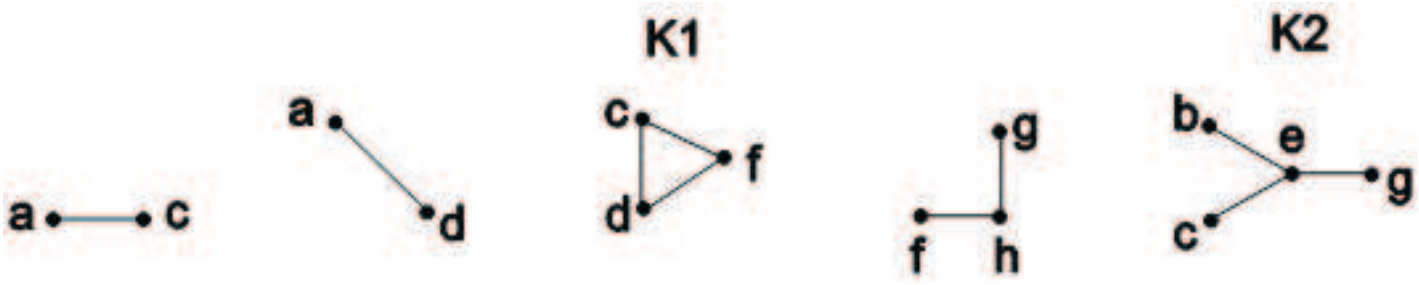


Figure: The local skeletons obtained from complete undirected subgraphs.

Bottom-up stage

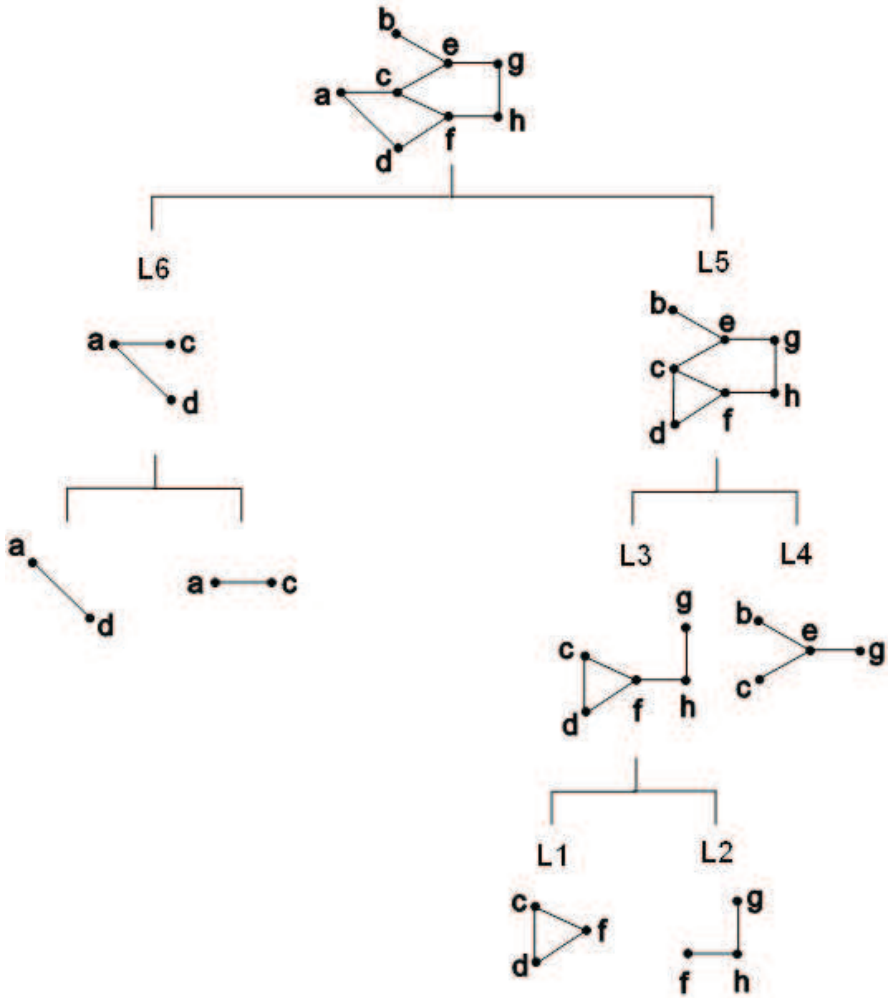


Figure: Combinations of local skeletons in Procedure CombineSubgraphs.

Bottom-up stage

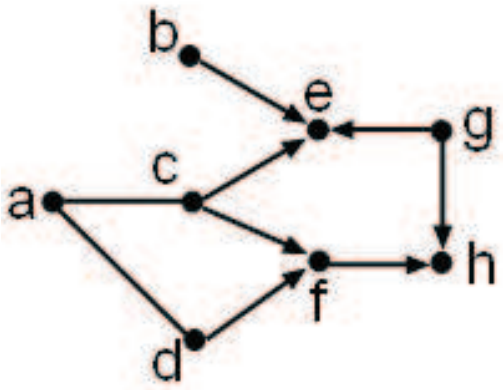


Figure: The constructed Markov equivalence class.

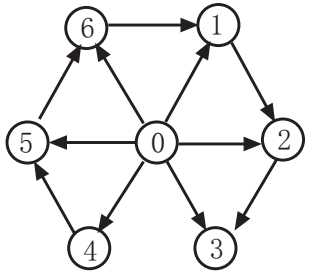
Outline

Active learning

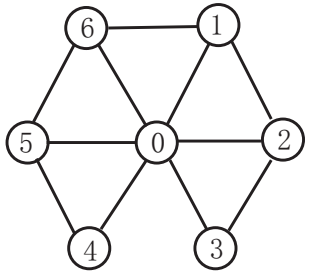
- Generally we cannot obtain causal relationships only using observational studies.
  There may be undirected edges which cannot be oriented by observational data.
- We propose an approach to determine causal directions by manipulation or intervention, called active learning.
- For $X_1 \to X_2$, manipulating cause X_1 changes $P(X_2)$ of effect; but manipulating effect X_2 cannot change $P(X_1)$ of cause.

Change an association network to a causal network

If data are generated from the unknown causal network



we can learn only an undirected association network



How to change it to a causal network?

We try to manipulate nodes as few as possible.

Active learning

We propose several manipulation approaches:
(He and G, 2008, JMLR)

- **Optimal batch manipulation**
  Find the minimum set of variables to be manipulated such that all edges can be oriented:

$$S_{\min} = \min\{S : \text{ manipulating } S \text{ can orient all edges}\}.$$

- **Random manipulation**
  Randomly select a variable to manipulate,
  Repeat manipulations until we can orient all edges.

Active Learning

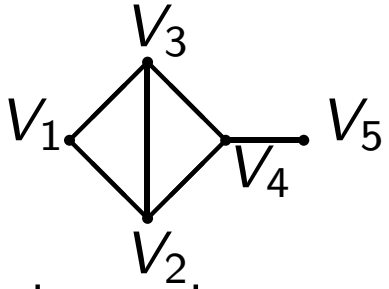
- **Optimal stepwise manipulation**
  - The MinMax criterion: manipulate a variable to minimize the maximum set of possible DAGs.
  - The maximum entropy criterion: manipulate a variable v to maximize the entropy

$$H_v = -\sum_{i=1}^{M} \frac{l_i}{L} \log \frac{l_i}{L}, \tag{1}$$

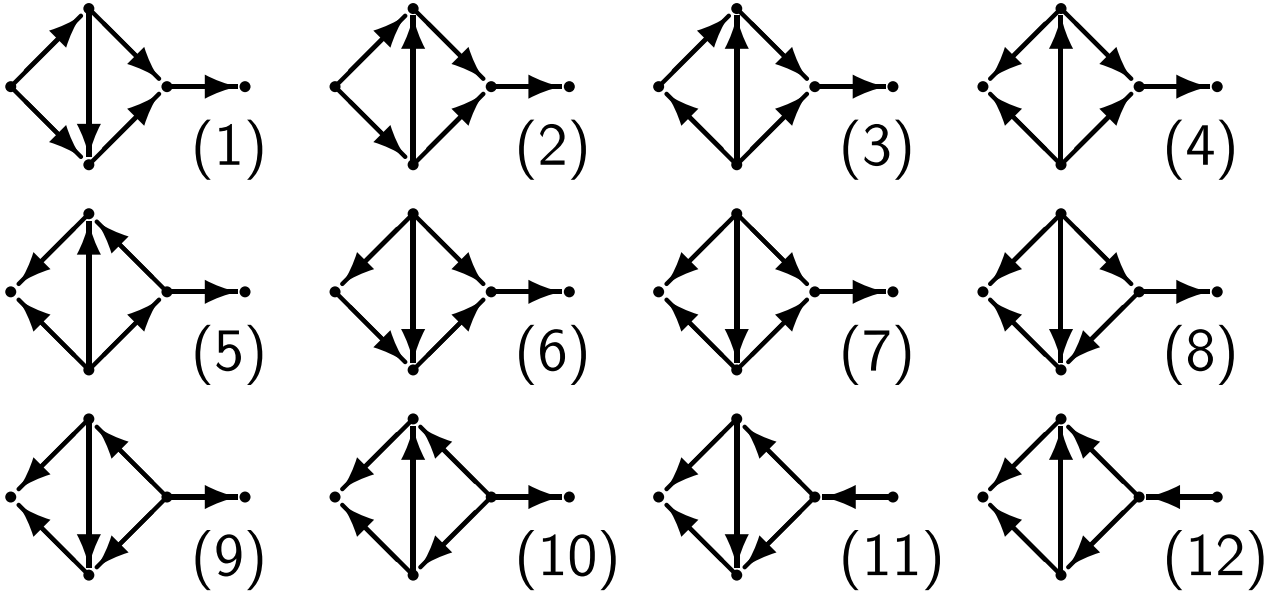
where M is the number of all possible orientation results obtained by manipulating a node v: $e(v)_1, \ldots, e(v)_M$;
l_i is the number of DAGs for ith orientation result $e(v)_i$;
$L = \sum_i l_i$.
That is, balance the sizes of DAG sets obtained by a manipulating.

Example of active learning

If we learnt the following Markov equivalent class $\bar{G}$ from data:



then the true causal network can be anyone of 12 DAGs



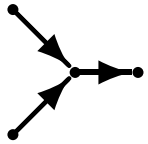
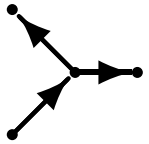
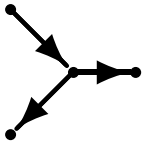
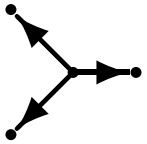
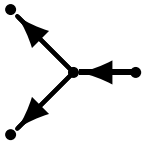
To orient $\bar{G}$, which variable should we manipulate first?

Example of manipulation

Table: Manipulate V_1

Orient	$V_2 \leftarrow V_1 \rightarrow V_3$	$V_2 \rightarrow V_1 \rightarrow V_3$	$V_2 \rightarrow V_1 \leftarrow V_3$	$V_2 \leftarrow V_1 \leftarrow V_3$
DAGs	$\{1, 2\}$	$\{3\}$	$\{4, 5, 7, 8, 9, 10, 11, 12\}$	$\{6\}$
l_i	2	1	8	1
Entropy is 0.9831 and maximum size is 8				

Table: Manipulate V_4

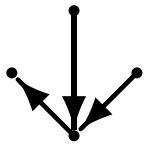
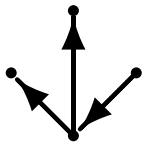
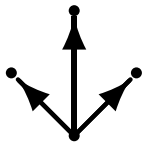
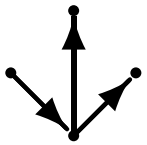
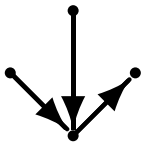
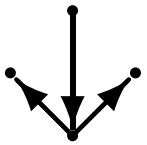
Orient					
DAGs	$\{1, 2, 3, 4, 6, 7\}$	$\{5\}$	$\{8\}$	$\{9, 10\}$	$\{11, 12\}$
l_i	6	1	1	2	2
Entropy is 1.3480 and maximum size is 6					

Example of manipulation

Table: Manipulate V_5

Orientation	$V_4 \rightarrow V_5$	$V_4 \leftarrow V_5$
DAGs	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$	$\{11, 12\}$
l_i	10	2
Entropy is 0.4506 and maximum size is 10		

Table: Manipulate V_2

Orient						
DAGs	$\{8, 9, 11\}$	$\{10, 12\}$	$\{3, 4, 5\}$	$\{2\}$	$\{1, 6\}$	$\{7\}$
l_i	3	2	3	1	2	1
Max Entropy is 1.7046 and Mini maximum size is 3						

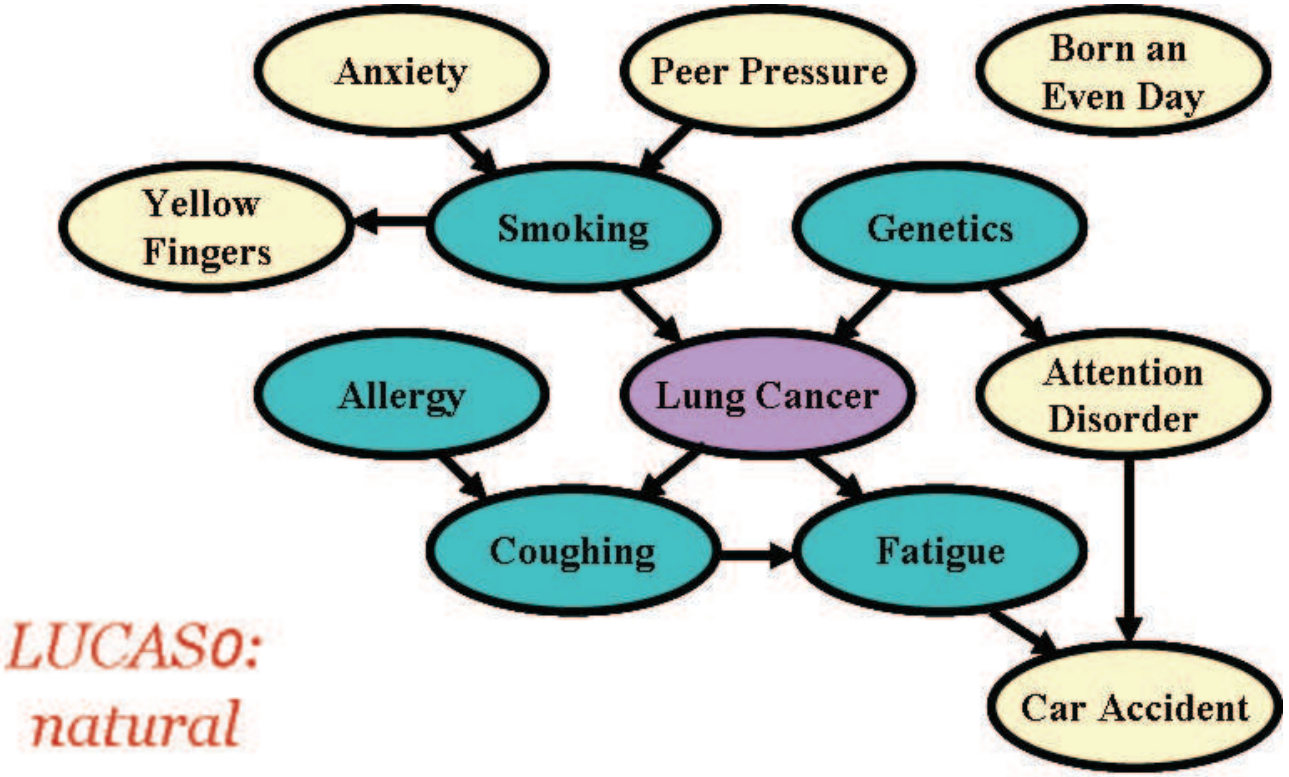
Outline

Causality

- Ordinary prediction approaches are based on association, which cannot do the prediction for the case with external interventions.
- For the case with the external interventions, we need to know what are the causes of a target variable.
- Commonly-used variable selection approaches cannot distinguish causes from effects.

Toy example by Guyon (2008)

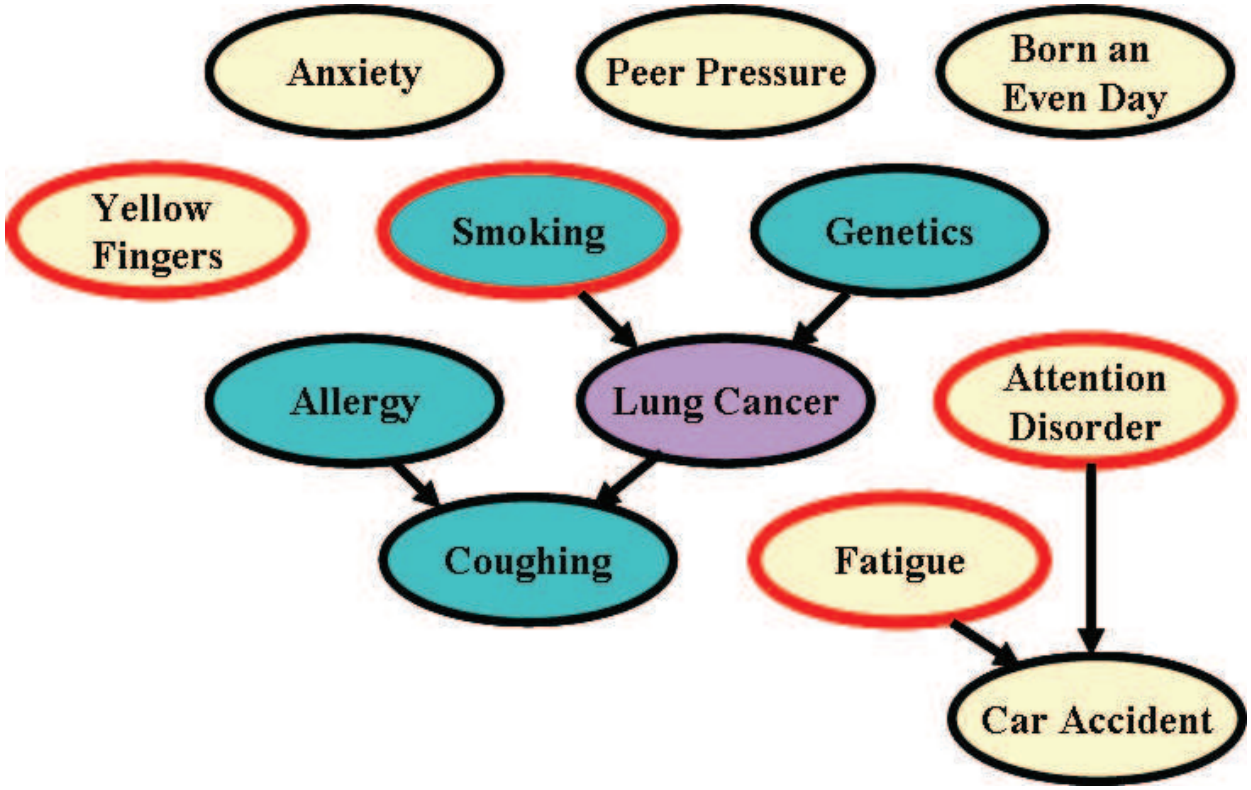
Guyon (2008) organized a causal challenge:
prediction for external intervention



Ordinary approaches cannot distinguish causes from effects,
and use the blue Markov blanket $MB(Y)$ to predict 'Lung Cancer',

Toy example by Guyon (2008)

If we manipulate these red nodes,
how to predict 'Lung Cancer'?



The manipulated Fatigue cannot be used for prediction.

Local learning of causal networks

- To find the causes of the target,
  one approach is to learn a whole causal network.
- But it is not necessary!
- We propose two approaches for local causal discovery:
  1. PCD-by-PCD algorithm (Zhou, Wang, Yin and G, 2010)
     (PCD: parents, children and descendants)
  2. MB-by-MB algorithm (Wang, Zhou, Zhao and G, 2014)
     (MB: Markov blanket)

Stepwise learning approaches

To discover the causes of the target T,

- first find all neighbours of T,
- then find the neighbours' neighbours of T,
  During finding neighbours, we can also find v-structures and orient the directions of some edges.
- Until we have determined all causes of T.

PCD-by-PCD approach

- **Initialization:**
  Set $WaitList = PCD(T)$.
  ($WaitList$ is the list of nodes whose PCDs will be found sequentially)

  Set $DoneList = \{T\}$.
  ($DoneList$ is the list of nodes whose PCDs have been found)

PCD-by-PCD algorithm(cont.)

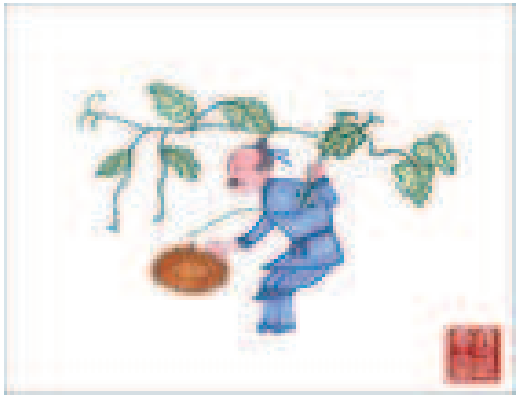
- **Repeat**
  - Take a node x from *WaitList*.
  - Find $PCD(x)$, put x into *DoneList*.
  - If $z \in PCD(x)$ and $x \in PCD(z)$, then create an edge (x, z).
  - Within *DoneList*, find v-structures $x \rightarrow z \leftarrow y$.
  - If new v-structures are found,
    orient other edges between nodes in *DoneList*.
  - Put $PCD(x)$ into *WaitList*
- **Until** (1) all edges connecting T are oriented,
  or (2) $WaitList = \emptyset$.

Example to illustrate PCD-by-PCD

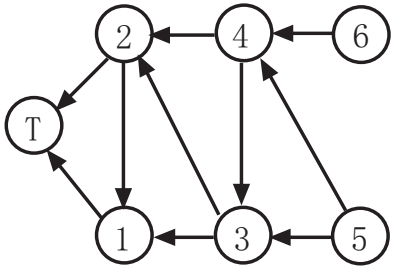
This algorithm can be demonstrated by two steps:

1 Trace to the root ; (寻根问底)

2 Follow the vine to get melon (顺藤摸瓜).



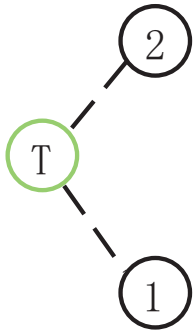
Suppose the unknown causal network:



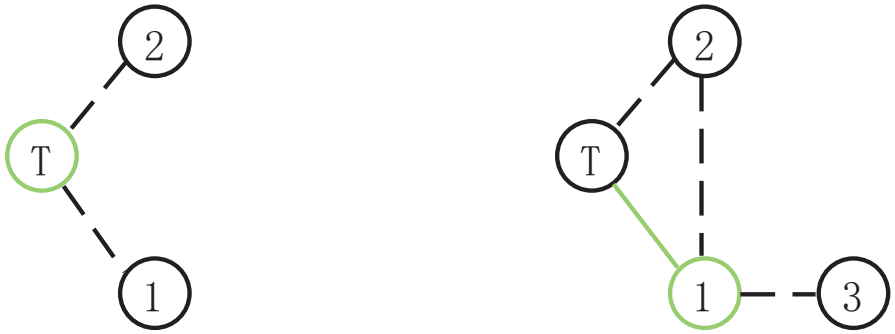
We want to find the direct causes of T.

Trace to the root (寻根问底)

- Find $PCD(T) = \{1, 2\}$.
- But we cannot determine whether there is an edge between T and 1 or an edge between T and 2 since nodes 1 and 2 may be descendants of T.
- Thus we use dash lines to denote the possible edges:



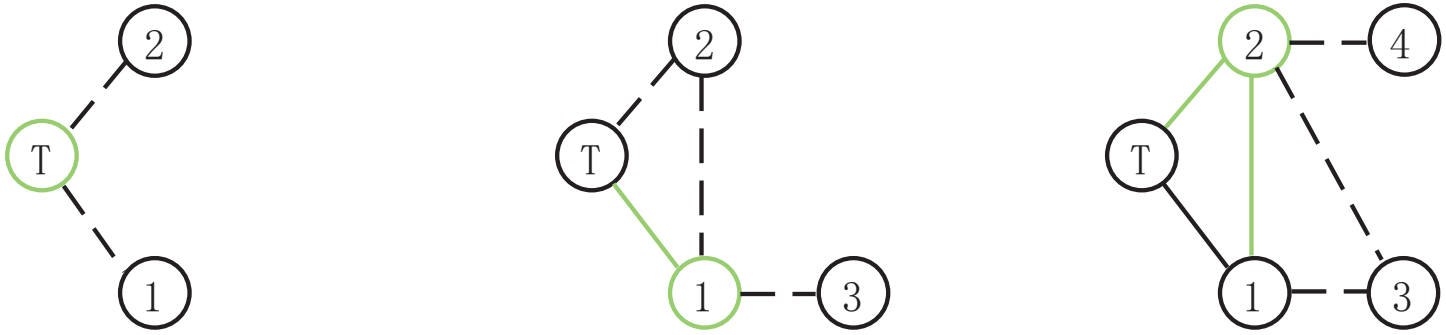
Trace to the root



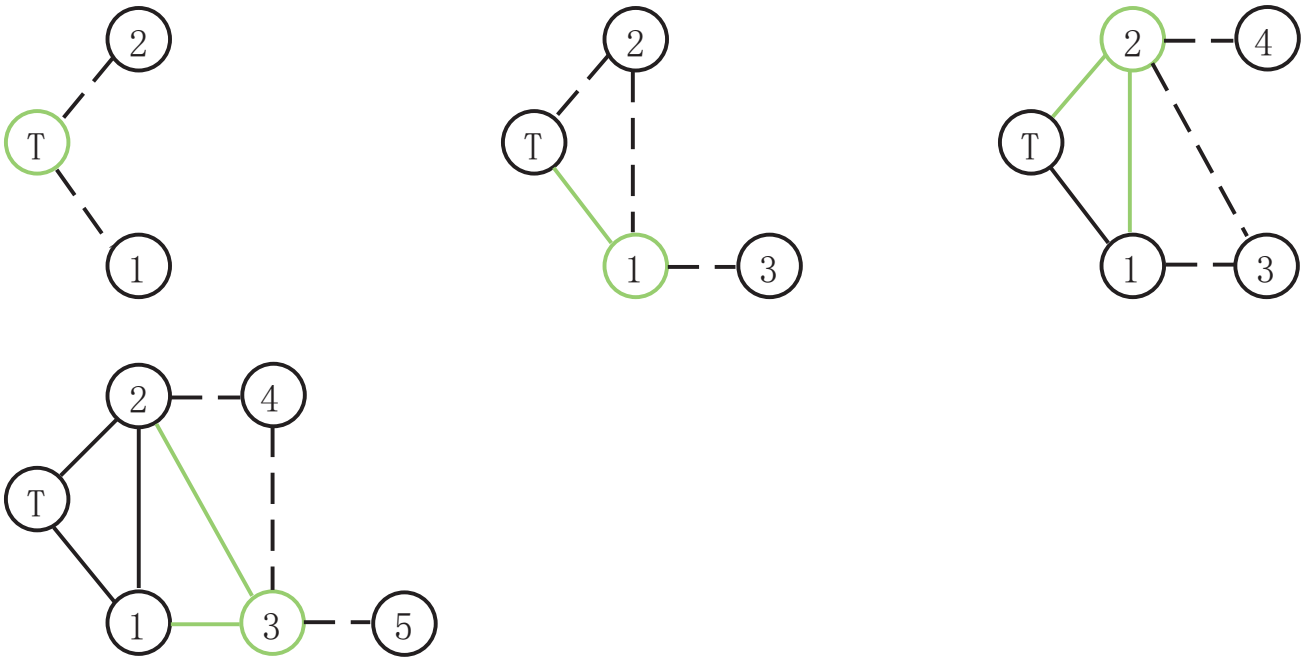
- Find $PCD(1) = \{T, 2, 3\}$.
- Because $1 \in PCD(T)$ and $T \in PCD(1)$, we can determine the edge between T and 1.
- Thus we change the dash line between T and 1 into a solid line.

Trace to the root

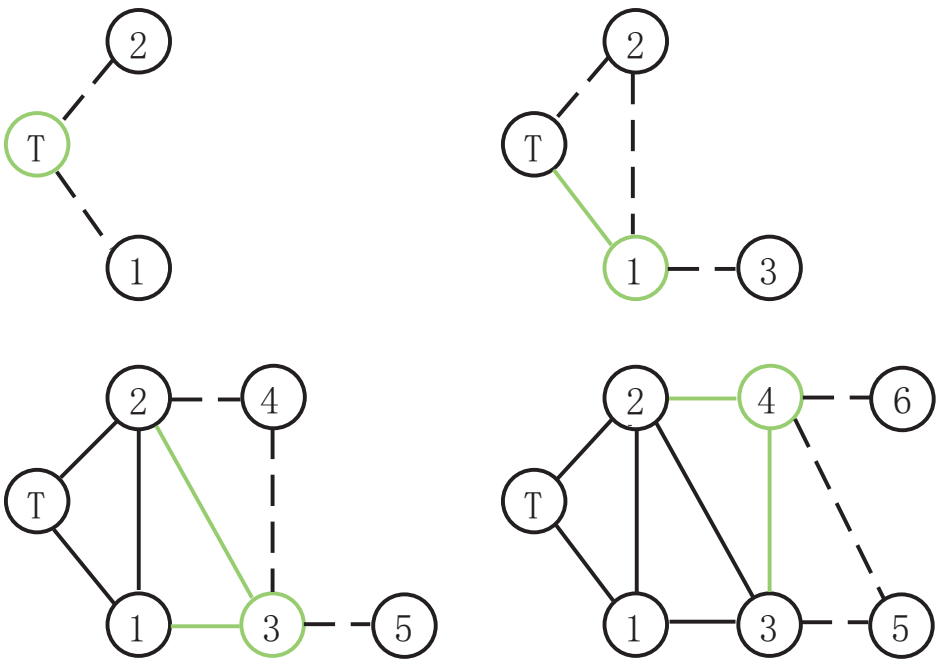
Similarly, find $PCD(2) = \{T, 1, 3, 4\}$.



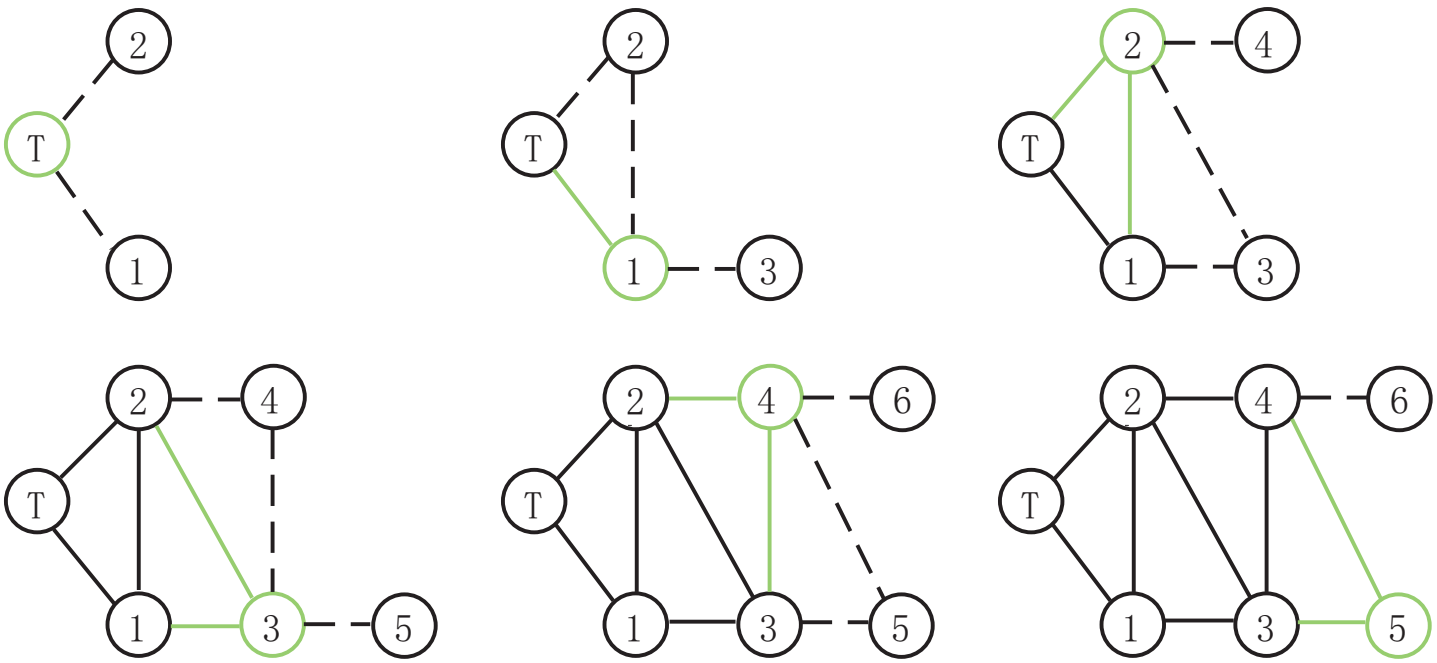
Trace to the root



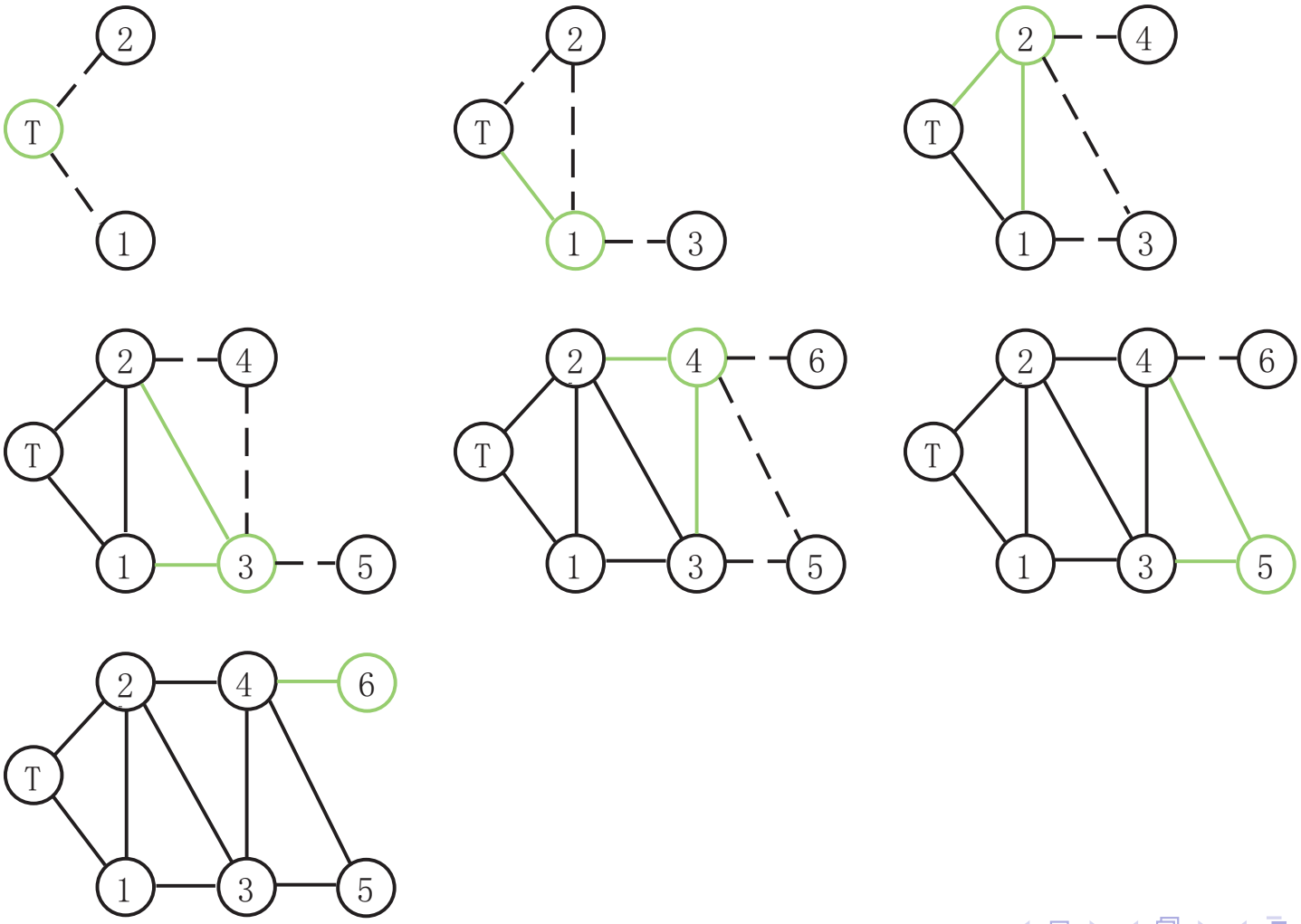
Trace to the root



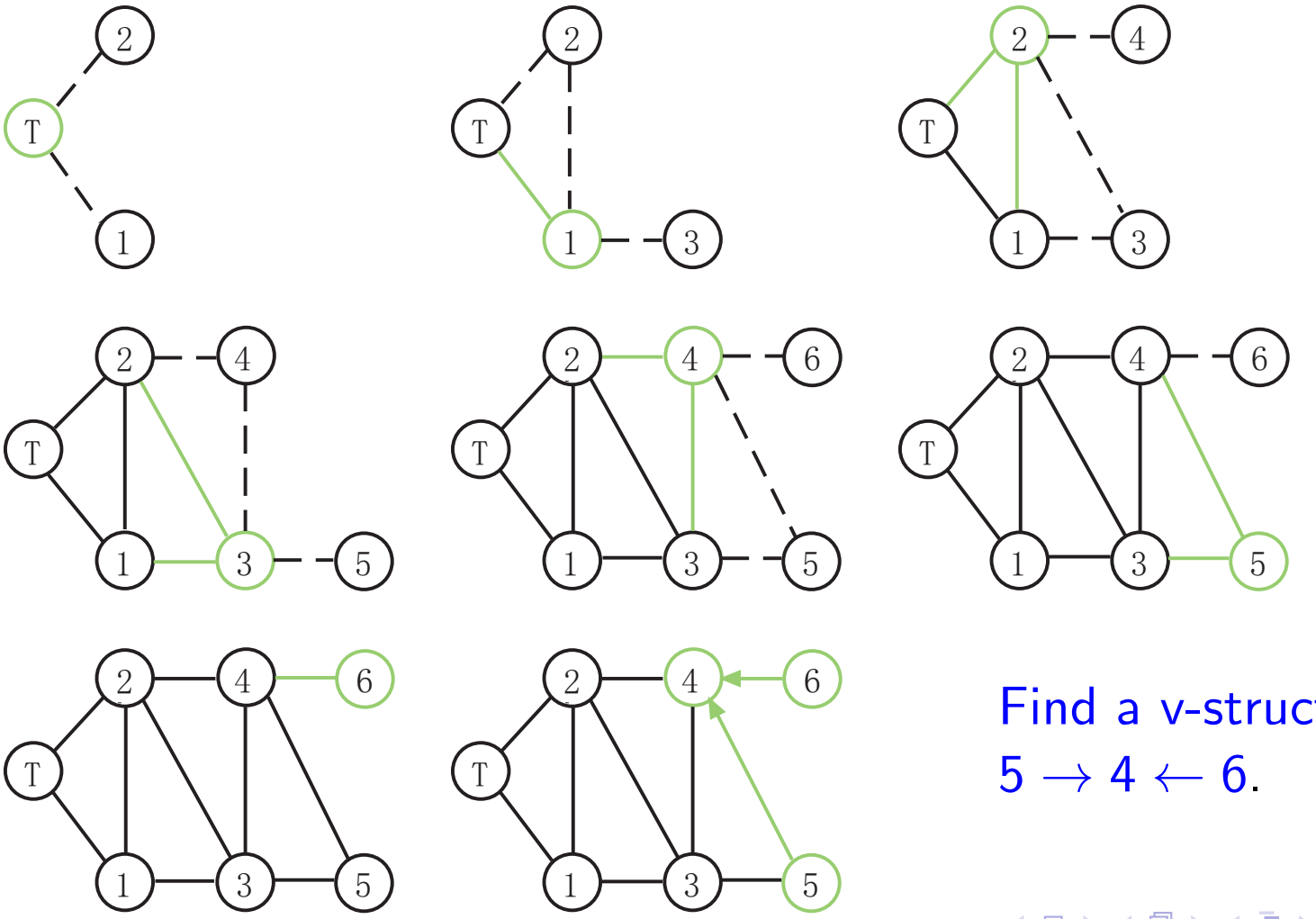
Trace to the root



Trace to the root



Trace to the root

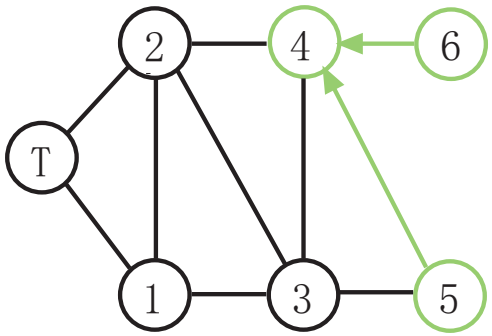


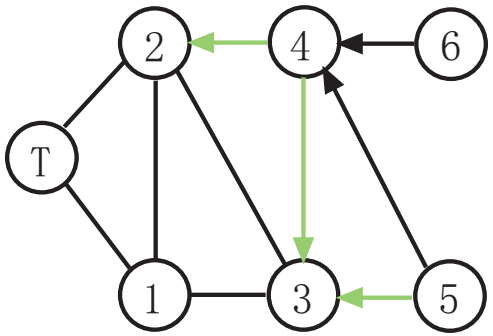
Find a v-structure
$5 \rightarrow 4 \leftarrow 6$.

Follow the vine to get the melon (顺藤摸瓜)

After finding the v-structure, we try to orient other edges:

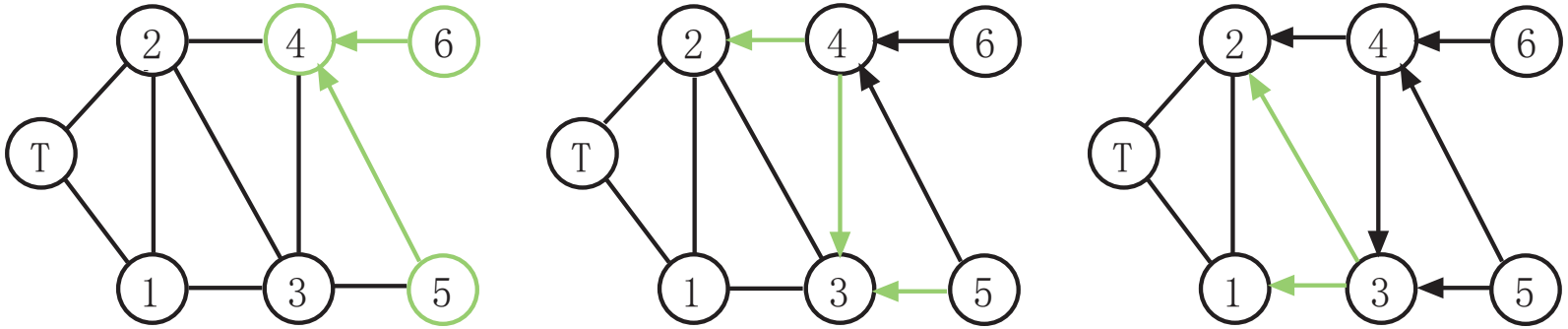
- $2 \leftarrow 4$, otherwise $2 \rightarrow 4 \leftarrow 6$ would make a new v-structure;
- $3 \leftarrow 4$, similar to above;
- $3 \leftarrow 5$, otherwise $3 \rightarrow 5$ would make a cycle.





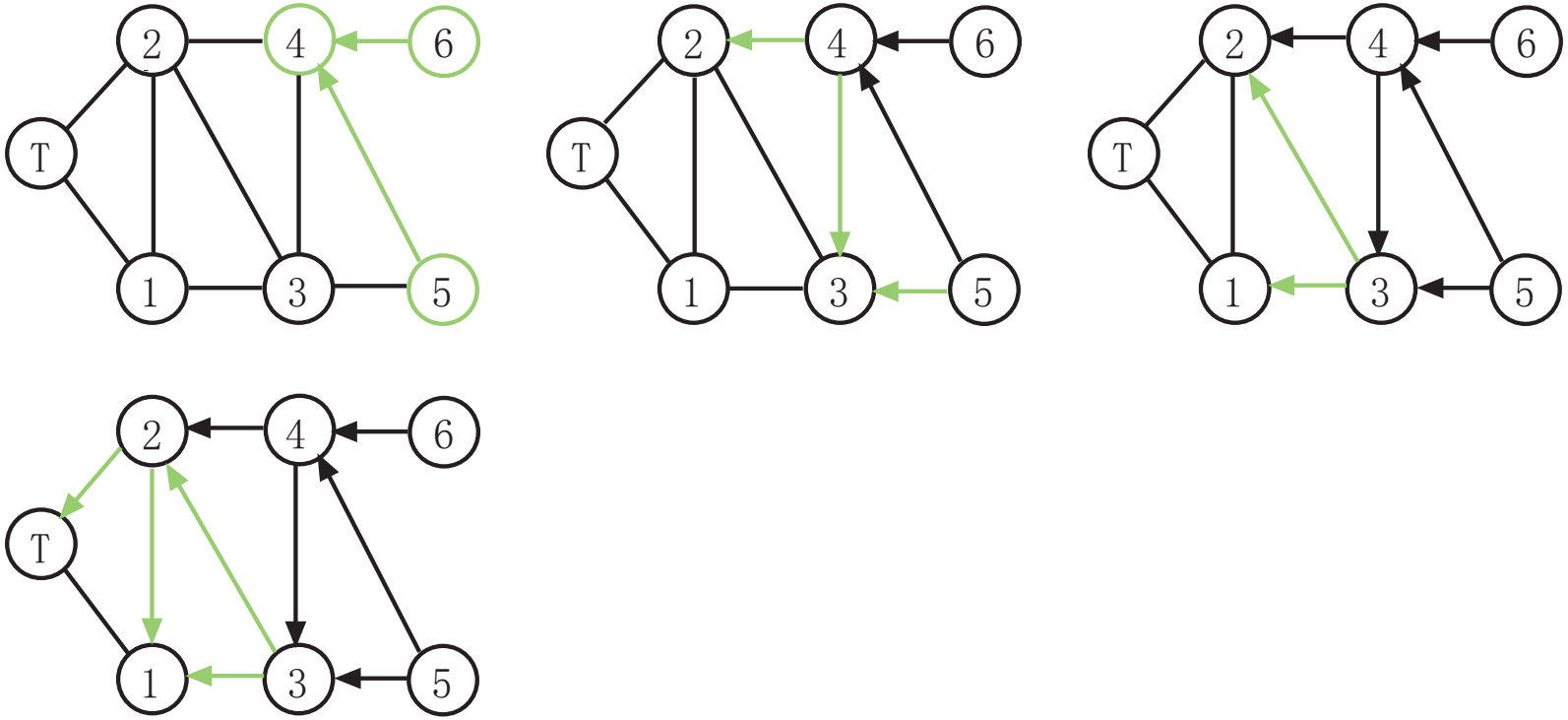
Follow the vine to get the melon

Similarly, we can orient all edges:



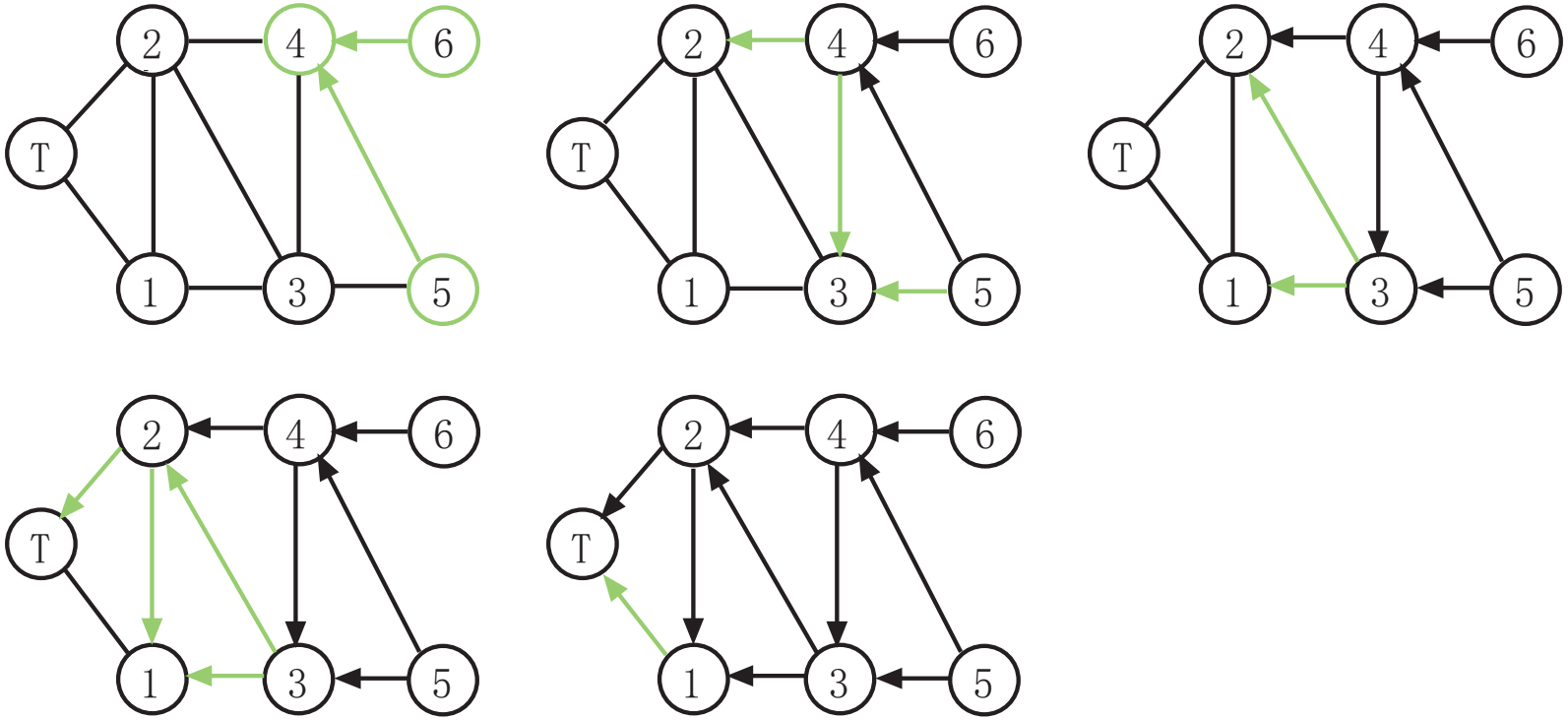
Follow the vine to get the melon

Similarly, we can orient all edges:



Follow the vine to get the melon

Similarly, we can orient all edges:



MB-by-MB algorithm

- There have been many approaches for variable selection, such as forward, stepwise and LASSO approaches, which can be used to find $MB(T)$:

$$T \perp\!\!\!\perp others | MB(T).$$

- Finding a MB of a node is easier than finding its PCD.
- Now we propose a local learning algorithm using variable selection.

MB-by-MB algorithm

The MB-by-MB Algorithm:

Input: a target T, observed data D.

1 **Initialization.**
WaitList $= T$; (WaitList keeps nodes whose MBs will be found)
$G = \emptyset$. (Initialize the graph around T)

2 **Repeat**
Take a node x from WaitList;
Find MB(x); Add MB(x) to WaitList.

3 Learn the local structure L_x over $MB(x) \cup \{x\}$.

4 Put the edges and the v-structures containing x in L_x to G.

5 Orient undirected edges in G.

6 **Until** (1) all edges connecting T are oriented or (2) WaitList$= \emptyset$.

Output: the local network G around T.

Example: ALARM

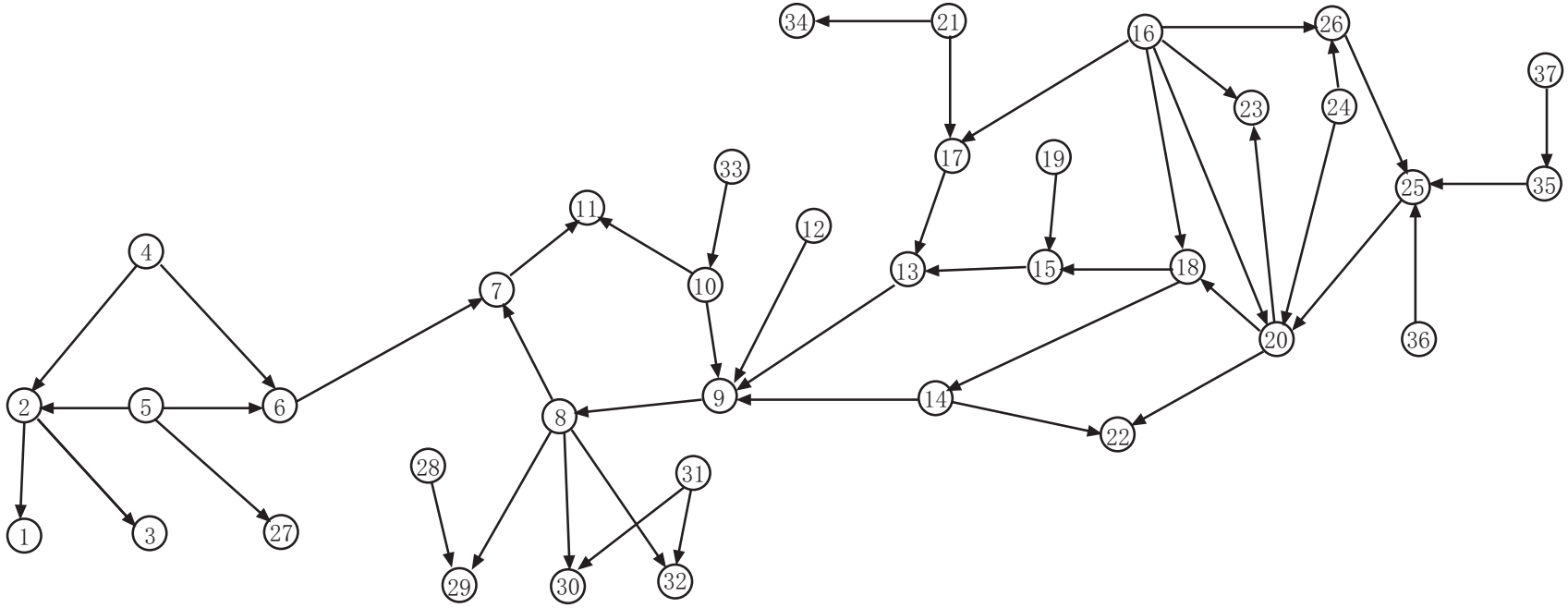
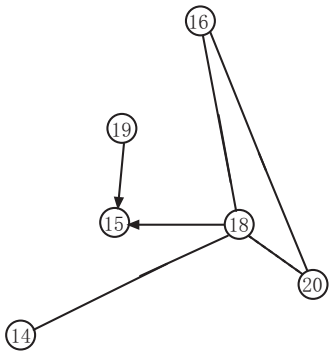


Figure: The ALARM network

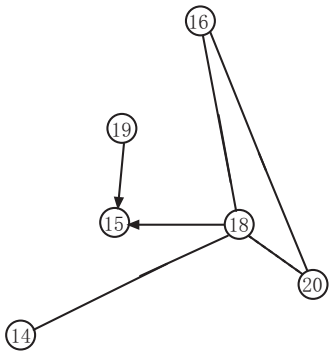
Suppose that node 18 is the target node.

Example: ALARM

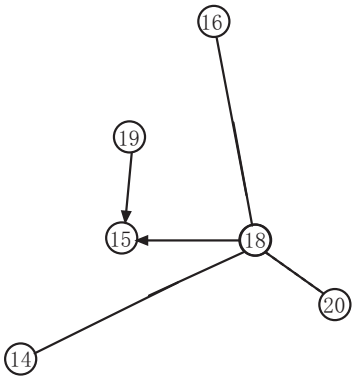


(a) $MB(18)$, L_{18}

Example: ALARM

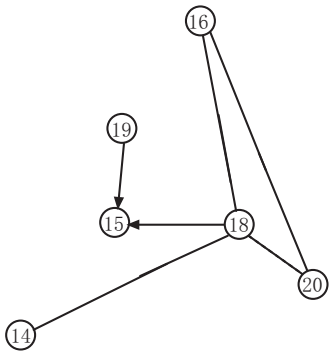


(a) $MB(18)$, L_{18}

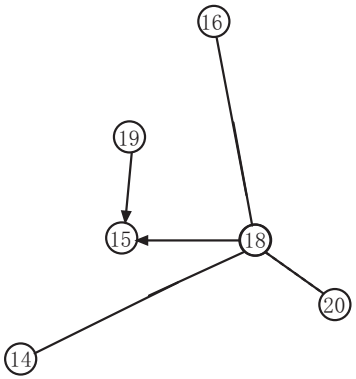


(b) Local G after learning L_{18}

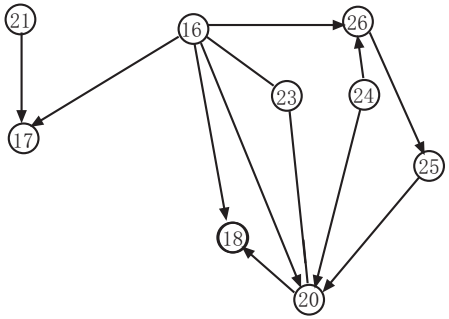
Example: ALARM



(a) $MB(18)$, L_{18}

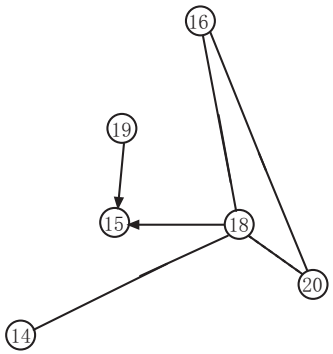


(b) Local G after learning L_{18}

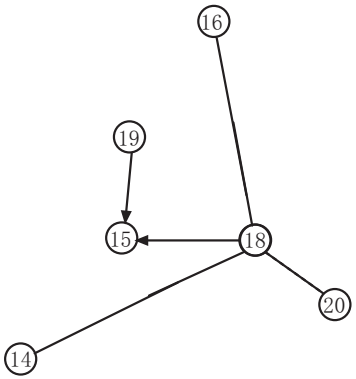


(c) $MB(16)$, L_{16}

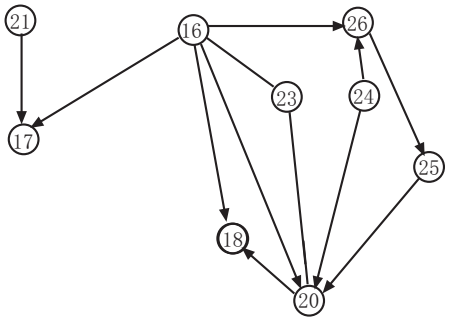
Example: ALARM



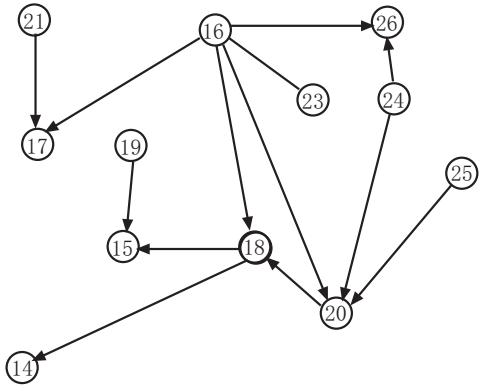
(a) $MB(18)$, L_{18}



(b) Local G after learning L_{18}



(c) $MB(16)$, L_{16}



(d) G around target 18

Figure: Sequential process to find causes and effects of node 18.

Summary

Topics	Approaches
Yule-Simpson paradox	Randomization, stratification, ...;
Surrogate paradox	Causation-based criteria, Association-based criteria for surrogates;
Decomposing learning	Learning from incomplete data, Recursive decomposition;
Active learning	Batch optimization, Step-wise optimizations;
Local learning	PCD-by-PCD algorithm, MB-by-MB algorithm;

Acknowledgements

Thank you!

These are joint works with my students:
Hua Chen, Ping He, Yangbo He, Chuan Ju, Changzhang Wang, Zhenguo Wu, Xianchao Xie, Jianxin Yin, You Zhou

References

Chen, H., Geng, Z. and Jia, J. (2007) Criteria for surrogate end points. J. Royal Statist. Soc. Ser. B 69, 919-932.

Deng, W., Geng, Z. and Li, H. (2013) Learning local directed acyclic graphs based on multivatriate time series data. Annals of Applied Statistics, 7, 1663-1683.

He, Y. and Geng, Z. (2008) Active learning of causal networks with intervention experiments and optimal designs. J. Machine Learning Research, 9, 2523-2547.

Ju, C. and Geng, Z. (2010) Criteria for surrogate endpoints based on causal distributions. J. Royal Statist. Soc. B 72, 129-142.

Wang, C. Z., Zhou, Y., Zhao, Q. and Geng, Z. (2014) Discovering and orienting the edges connected to a arget variable in a DAG via a sequential local learning approach. Comput. Statist. & Data Analy., 77,252-266.

Wu Z. G., He, P. and Geng, Z. (2011) Sufficient conditions for concluding surrogacy based on observed data. Statist. Medicine, 30, 2422-2434..

Xie, X. and Geng, Z. (2008) A recursive method for structural learning of directed acyclic graphs. *J Machine Learning Research*, 9, 459-483.

Xie, X., Geng, Z. and Zhao, Q. (2006) Decomposition of structural learning about directed acyclic graphs. Artificial Intelligence, 170, 422-439.