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

1 Causal Effect Evaluation
   - Yule-Simpson paradox
   - Causal effects
   - Surrogate and surrogate paradox

2 Causal Network Learning
Outline

1. Causal Effect Evaluation
   - Yule-Simpson paradox
   - Causal effects
   - Surrogate and surrogate paradox

2. Causal Network Learning
"Human can be compared to a frog at the bottom of a well"

Frog’s sight $\Rightarrow$ Can the frog make a correct inference about the universe from its sight?

Frog $\Rightarrow$
# Yule-Simpson Paradox (Yule, 1900; Simpson, 1951)

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Non-smoking</td>
<td>80</td>
<td>120</td>
<td>200</td>
</tr>
</tbody>
</table>

\[ RD = \frac{100}{200} - \frac{80}{200} = 0.10 \]

<table>
<thead>
<tr>
<th></th>
<th>Male (Gene=+)</th>
<th>Female (Gene=−)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cancer</td>
<td>Control</td>
</tr>
<tr>
<td>Smoking</td>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>Non-smok</td>
<td>35</td>
<td>15</td>
</tr>
</tbody>
</table>

\[ RD_M = \frac{90}{150} - \frac{35}{50} = -0.10 \]
\[ RD_F = \frac{10}{50} - \frac{45}{150} = -0.10 \]

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

1. Causal Effect Evaluation
   - Yule-Simpson paradox
   - Causal effects
   - Surrogate and surrogate paradox

2. Causal Network Learning
Definitions of Causal Effects (Neyman, 1923; Rubin, 1974)

- For an individual $i$,
  \[ Y_1(i): \text{potential outcome if treatment } T \text{ were 1 (Smoking)}, \]
  \[ Y_0(i): \text{potential if treatment } T \text{ were 0 (Non-smoking)}, \]

- 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):
  \[ ACE(T \rightarrow Y) = E(Y_1 - Y_0) = E(Y_1) - E(Y_0). \]
Causal effect ≠ Association measure

- Generally, ACE is not identifiable.

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

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

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

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 \( (Y_1, Y_0) \perp T | X \), where \( X \) is a sufficient confounder set.
  - If \( X \) is observed, then
    \[
    ACE(T \rightarrow Y) = \sum_x ACE(T \rightarrow Y|x)P(x).
    \]
  - No Yule-Simpson paradox for ACE:
    \[
    ACE(T \rightarrow Y|x) > 0, \forall x \iff ACE(T \rightarrow Y) > 0.
    \]
  - 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) \( Z \) (\( Z \perp T \) and \( Z \perp X \)), to estimate ACE.
Outline

1 Causal Effect Evaluation
   - Yule-Simpson paradox
   - Causal effects
   - Surrogate and surrogate paradox

2 Causal Network Learning
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**.
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 $Y$, $Y \perp T \mid S$.

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

   $$S_{T=1}(u) = S_{T=0}(u) \quad \Rightarrow \quad p(Y_{T=0}) = p(Y_{T=1}), \text{ for these } u.$$
Criteria for Surrogates

- **The strong surrogate criterion** (Lauritzen, 2004):

\[
\begin{align*}
T & \rightarrow S \\
S & \rightarrow Y
\end{align*}
\]

where \( U \) is an unobserved variable.

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

$$ACE(T \rightarrow S) = +$$

\[ T \rightarrow S \rightarrow Y \]
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 $Y$.

\[
\text{ACE}(T \rightarrow S) = + \quad \text{ACE}(S \rightarrow Y) = +
\]

\[T \xrightarrow{} S \xrightarrow{} Y\]
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 $Y$.

\[
\text{ACE}(T \rightarrow S) = + \quad \text{ACE}(S \rightarrow Y) = + \quad \text{ACE}(T \rightarrow Y) = -
\]
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 $Y$.

\[
ACE(T \rightarrow S) = + \quad ACE(S \rightarrow Y) = + \quad ACE(T \rightarrow Y) = -
\]

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

- 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 intermediary \(S\), 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)

<table>
<thead>
<tr>
<th>Group</th>
<th>No.</th>
<th>$S_{T=0}$</th>
<th>$S_{T=1}$</th>
<th>$Y_{S=0} &lt; Y_{S=1}$</th>
<th>$Y_{T=0}$</th>
<th>$Y_{T=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
ACE(T \rightarrow S) = \frac{40 + 20}{100} - \frac{20 + 20}{100} = \frac{20}{100} > 0, \]

but

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

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

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

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

- **Goal:** Without observing $Y$, but observing $S$ instead, we want to predict the sign ($+, -, 0$) of $DCE[T \rightarrow (Y > y)]$ using the sign of $DCE[T \rightarrow (S > s)]$.

- To avoid the surrogate paradox, we give different conditions, some are based on **associations**, and some are based on **causations**.
Theorem 1. (Ju and G, JRSS B, 2010)
Assume that the causal network is true: without $T \rightarrow Y$

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

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

\[ DCE[T \rightarrow (S > s)] \geq 0 \iff DCE[T \rightarrow (Y > y)] \geq 0 \]

- The conditions are testable if \( Y \) is observed in a validation study.
- But the reverse ‘\( \leftarrow \)’ is not true.
Theorem 3. If

1. Prentice’s criterion $Y \perp\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\! T \mid S$,  
2. $P(Y > y \mid s)$ increases in $s$ and  
3. $S$ is from an exponential family conditional on $T$,

then

\[ \text{Sign}[\text{ACE}(T \rightarrow S)] = \text{Sign}[\text{DCE}(T \rightarrow S)] \]
\[ = \text{Sign}[\text{ACE}(T \rightarrow Y)] = \text{Sign}[\text{DCE}(T \rightarrow Y)], \]

where \text{Sign} means ‘$= 0’$, ‘$> 0’$ or ‘$< 0’$.\]
Summary of criteria for surrogates

- The **principal surrogate** and the **strong surrogate**: only
  \[ CE(T \rightarrow S) = 0 \iff CE(T \rightarrow Y) = 0. \]

- The **monotonicity**: further
  \[ CE(T \rightarrow S) \geq (\leq) 0 \iff CE(T \rightarrow Y) \geq (\leq) 0. \]

- Prentice’s criterion and **S** from the exponential family: equivalence relationships
  \[ CE(T \rightarrow S) > (<, =) 0 \iff CE(T \rightarrow Y) > (<, =) 0. \]
Outline

1. Causal Effect Evaluation

2. Causal Network Learning
   - Decomposing learning
   - Active learning
   - Local learning
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

1. Causal Effect Evaluation

2. Causal Network Learning
   - Decomposing learning
   - Active learning
   - Local learning
We discuss how blind men can discover an elephant:

(Xie, G and Zhao, 2006, Artificial Intelligence)
The decomposing approach:

- Three experts in different areas observed different variable sets.
- We obtained 3 incomplete data sets of the variable sets.
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:
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:
Combining subgraphs and orienting other undirected edges, we obtain the **Markov equivalence class**:
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 DecomLearning \((K, \bar{L}_K)\)

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

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

1. Causal Effect Evaluation

2. Causal Network Learning
   - Decomposing learning
   - Active learning
   - Local 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 $X_1 \rightarrow X_2$, 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

![Diagram of a causal network](image1)

we can learn only an undirected association network

![Diagram of an association network](image2)

How to change it to a causal network?
We try to manipulate nodes as few as possible.
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:

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

- **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

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

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 \( i \)th 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:

$$
\begin{align*}
V_1 &\xrightarrow{\quad} V_3 \\
V_2 &\xrightarrow{\quad} V_4 \\
V_4 &\xrightarrow{\quad} V_5
\end{align*}
$$

then the true causal network can be anyone of 12 DAGs:

1. \[ (1) \]
2. \[ (2) \]
3. \[ (3) \]
4. \[ (4) \]
5. \[ (5) \]
6. \[ (6) \]
7. \[ (7) \]
8. \[ (8) \]
9. \[ (9) \]
10. \[ (10) \]
11. \[ (11) \]
12. \[ (12) \]

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

**Table: Manipulate \( V_1 \)**

<table>
<thead>
<tr>
<th>Orient DAGs</th>
<th>( V_2 \leftarrow V_1 \rightarrow V_3 )</th>
<th>( V_2 \rightarrow V_1 \rightarrow V_3 )</th>
<th>( V_2 \rightarrow V_1 \leftarrow V_3 )</th>
<th>( V_2 \leftarrow V_1 \leftarrow V_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_i )</td>
<td>{1, 2}</td>
<td>{3}</td>
<td>{4, 5, 7, 8, 9, 10, 11, 12}</td>
<td>{6}</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.9831</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table: Manipulate \( V_4 \)**

<table>
<thead>
<tr>
<th>Orient DAGs</th>
<th>{1, 2, 3, 4, 6, 7}</th>
<th>{5}</th>
<th>{8}</th>
<th>{9, 10}</th>
<th>{11, 12}</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_i )</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Entropy** is 0.9831 and maximum size is 8

**Entropy** is 1.3480 and maximum size is 6
Example of manipulation

**Table: Manipulate V₅**

<table>
<thead>
<tr>
<th>Orientation</th>
<th>V₄ → V₅</th>
<th>V₄ ← V₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAGs</td>
<td>{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}</td>
<td>{11, 12}</td>
</tr>
<tr>
<td>lᵢ</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.4506</td>
<td></td>
</tr>
<tr>
<td>Maximum size</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Table: Manipulate V₂**

<table>
<thead>
<tr>
<th>Orient</th>
<th>DAGs</th>
<th>lᵢ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{8, 9, 11}</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>{10, 12}</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>{3, 4, 5}</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>{2}</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>{1, 6}</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>{7}</td>
<td>1</td>
</tr>
</tbody>
</table>

Max Entropy is 1.7046 and Mini maximum size is 3
Outline

1. Causal Effect Evaluation

2. Causal Network Learning
   - Decomposing learning
   - Active learning
   - Local learning
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.

Ordinary approaches cannot distinguish causes from effects, and use the blue Markov blanket $MB(Y)$ to predict ‘Lung Cancer’.
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:
     (PCD: parents, children and descendants)
  2. MB-by-MB algorithm (Wang, Zhou, Zhao and G, 2014)
     (MB: Markov blanket)
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)
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 $T$. 
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:
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.
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.
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:
Similarly, we can orient all edges:

```
Similarly, we can orient all edges:
```

![Diagram showing oriented edges](image)

Follow the vine to get the melon
Similarly, we can orient all edges:

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

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

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.**
   - $\text{WaitList} = T$; ($\text{WaitList}$ keeps nodes whose MBs will be found)
   - $G = \emptyset$. (Initialize the graph around $T$)

2. **Repeat**
   - Take a node $x$ from $\text{WaitList}$;
   - Find $MB(x)$; Add $MB(x)$ to $\text{WaitList}$.

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) $\text{WaitList} = \emptyset$.

**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.
<table>
<thead>
<tr>
<th>Topics</th>
<th>Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yule-Simpson paradox</td>
<td>Randomization, stratification, ...;</td>
</tr>
<tr>
<td>Surrogate paradox</td>
<td>Causation-based criteria, Association-based criteria for surrogates;</td>
</tr>
<tr>
<td>Decomposing learning</td>
<td>Learning from incomplete data, Recursive decomposition;</td>
</tr>
<tr>
<td>Active learning</td>
<td>Batch optimization, Step-wise optimizations;</td>
</tr>
<tr>
<td>Local learning</td>
<td>PCD-by-PCD algorithm, MB-by-MB algorithm;</td>
</tr>
</tbody>
</table>
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


