

# Confounder selection via iterative graph expansion

---

F. Richard Guo

31 Oct, 2023

Online Causal Inference Seminar

Statistical Laboratory, University of Cambridge



Qingyuan Zhao

# Outline

Introduction

Demo time

Theory

# Introduction

## Confounder selection

In observational studies, the single most widely used to strategy to control for confounding is through confounder / covariate **adjustment**.

► Suppose  $X$  is a (point) treatment and  $Y$  is an outcome.

☞ Adjust for **a set of covariates  $S$**  such that

$Y(x) \perp\!\!\!\perp X \mid S$ , for every treatment level  $x$ . ► conditional exchangeability

☞ Then, under positivity,

$$p(Y(x) \mid X = x, S) = p(Y \mid X = x, S).$$

☞ **Confounder selection**: How to select  $S$ ?

# Objectives

1. **Primary:** Finding a set of **observed** covariates  $S$  that satisfies conditional exchangeability  $Y(x) \perp\!\!\!\perp X \mid S$ .
  - ▶ Such a set is called a **sufficient adjustment set**.
2. **Secondary:** When there are more than one sufficient adjustment sets, choose one among them to optimize some criterion, such as
  - efficiency
  - cardinality
  - cost
  - ...

👉 We only focus on the **primary objective** in this talk.

- ▶ See also Guo, Lundborg, and Zhao (2022) for a recent survey.

## A solved problem?

👉 Suppose we want to find a sufficient adjustment set in the causal model<sup>1</sup> represented by a DAG (▶ or ADMG)  $\mathcal{G}$  over vertex set  $V$ .

**Back-door Criterion** (Pearl, 1993)  $S \subseteq V \setminus \{X, Y\}$  is a sufficient adjustment set if

1.  $S$  contains no descendant of  $X$ ,
2. there is no ' $X \leftarrow \dots$ ' path between  $X$  and  $Y$  that is  $m$ -connected given  $S$ .

**Completeness** (Shpitser, VanderWeele, and Robins, 2010) If  $S$  is a sufficient adjustment set under the causal model  $\mathcal{G}$ , then  $S \setminus \text{de}(X)$  satisfies the backdoor criterion.

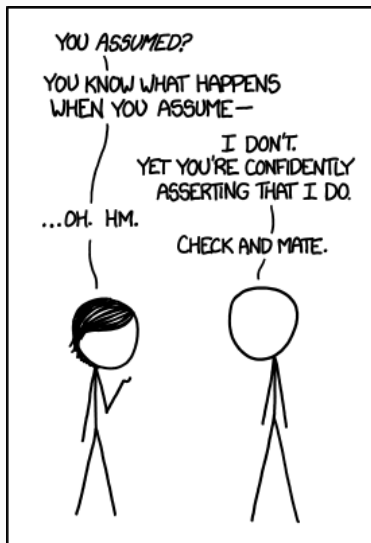
▶ WLOG, we only consider adjustment sets that contain **no descendants** of  $X$ , then

$S$  is a sufficient adjustment set  $\iff S$  meets the back-door criterion.

---

<sup>1</sup>Formally, the FFRCISTG/SWIG model, which is a supermodel of NPSEM-IE.

- ▶ “Suppose  $\mathcal{G}$  is the causal DAG/ADMG ...”





# Critiques on the 'known $\mathcal{G}$ ' stance

## 1. **Impractical**

- Do not know the full causal structure/mechanism
- Even if we know it, can we readily draw it?
  - Where is the boundary?
  - Can you draw the floor plan of your home?
- Ask a domain expert to draw it
  - Tools and protocols (Shrier and Platt, 2008; Haber et al., 2022) are developed, but still challenging.

## 2. **Unnecessary**: Back-door criterion only concerns **partial knowledge** about $\mathcal{G}$ .

### ▶ **Two questions**

- (i) Can we represent this partial knowledge in a modular format? (▶ representation)
- (ii) How do we elicit this partial knowledge? (▶ procedure design)

**Disjunctive criterion** (VanderWeele and Shpitser, 2011):

$$S := (\text{an}_g(X) \cup \text{an}_g(Y)) \cap \{\text{observed pre-treatment covariates}\}.$$

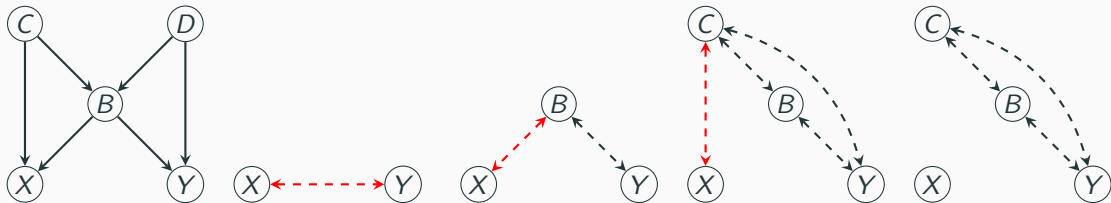
▶ See also VanderWeele (2019) for its variation.

☞ Such  $S$  is a **sufficient adjustment set** whenever  $\{\text{observed pre-treatment covariates}\}$  contains **any** sufficient adjustment set.

☞ Most useful when

- (1) data is already collected,
- (2) structural knowledge is scarce.

## Our approach in a nutshell



- (i) Can we represent this partial knowledge in a modular format? (► representation)
- 👉 Graph with bidirected edges: ' $\leftarrow - \rightarrow$ ' denotes a (potential) **confounding arc**
- (ii) How do we elicit this partial knowledge? (► procedure design)
- 👉 Select a confounding arc and introduce new variables (► **primary adjustment set**) to control it. (► knowledge about **common causes** and **mediators**)
  - 👉 Remove the old arc and adds new arcs.
  - 👉 Iterate until X and Y are disconnected — a sufficient adjustment set is found!
  - (► inverse of latent projection)

**Demo time**



<https://ricguo.shinyapps.io/InteractiveConfSel/>

Example

# Features of the procedure

- **Confounding arcs**

$\leftarrow - \rightarrow$  : potential confounding arc

$\longleftrightarrow$  : uncontrollable confounding arc (▶ no primary adjustment set exists)

- To control  $\leftarrow - \rightarrow$ , knowledge is elicited to find its **primary adjustment set**.

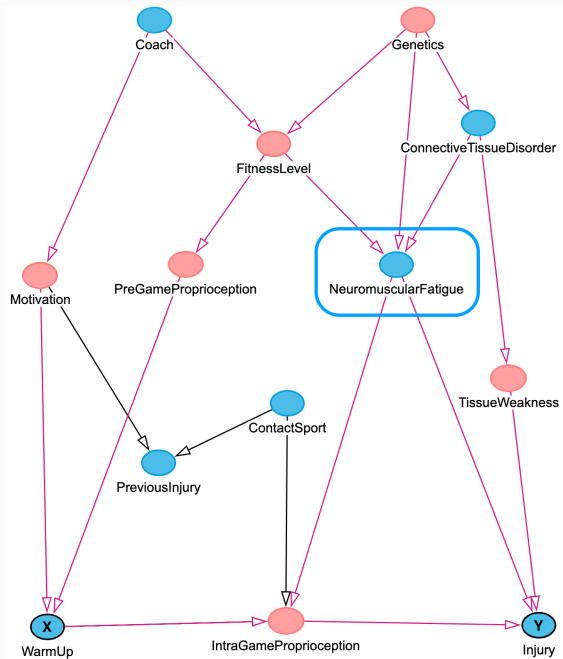
- Does not need pre-specification of the full graph.

- User answers questions about **common causes** and **mediators**. (▶ Local structures.)

- **Economical** queries: causal relations between the observed confounders are irrelevant and never asked about!

- User's familiarity with causal graphical models is **not** a prerequisite.

- ▶ Need not be aware of **collider bias** — it is taken care of!



# Features of the procedure

- **Confounding arcs**

- $\leftarrow - \rightarrow$ : potential confounding arc;

- $\longleftrightarrow$ : uncontrollable confounding arc

- To control  $\leftarrow - \rightarrow$ , knowledge is elicited to find its **primary adjustment set**.

- Does not need pre-specification of the full graph.

- User answers questions about **common causes** and **mediators**. (▶ Local structures.)

- **Economical** queries: causal relations between the observed confounders are irrelevant and never asked about!

- User's familiarity with causal graphical models is **not** a prerequisite.

- ▶ Need not be aware of **collider bias** — it is taken care of!

- Procedure is terminated when  $X$  and  $Y$  are

- ▶ disconnected by  $\leftarrow - \rightarrow / \longleftrightarrow$  edges: a sufficient adjustment set is found;

- ▶ connected by  $\longleftrightarrow$  edges: no sufficient adjustment set can be found.



**Theory**

# Setting

Let  $\mathcal{G}$  be an underlying acyclic directed mixed graph (ADMG) over vertex set  $V$  that defines our causal model.

- ▶ An ADMG is a graph with directed ( $\rightarrow$ ) and bidirected edges ( $\leftrightarrow$ ) that has no directed cycle.
  - ' $\rightarrow$ ' represents a (direct) causal effect.
  - ' $\leftrightarrow$ ' represents the existence of a latent common cause, i.e., *endogeneity*.

**m-connection/separation** (Richardson, 2003) is a direct extension of d-connection/separation:

- ▶ A path from  $A$  to  $B$  is m-connected given  $C$  if every non-collider on the path is not in  $C$ , and every collider on the path is in  $C$  or has a descendant in  $C$ .

$$A \perp\!\!\!\perp_m B \mid C \iff \nexists \text{ m-connected path between } A \text{ and } B \text{ given } C$$

$$A \not\perp\!\!\!\perp_m B \mid C \iff \exists \text{ m-connected path between } A \text{ and } B \text{ given } C$$

 Graph  $\mathcal{G}$  is **unknown**. But we can make **repeated queries** about certain structures in  $\mathcal{G}$ .

## Notation: shapes of paths

**Arc** '⋈' = a sequence of adjacent edges with no colliders

☞ Directed arc  $A \rightsquigarrow B$ :  $A \rightarrow \dots \rightarrow B$

☞ **Confounding arc**  $A \leftrightarrow B$ :

$A \leftarrow \dots \rightarrow B$  or  $A \leftarrow \dots \leftarrow \circ \leftrightarrow B$  or  $A \leftrightarrow \circ \rightarrow \dots \rightarrow B$  or  $A \leftarrow \dots \leftrightarrow \dots \rightarrow B$

**Half-arrow** = either an endpoint tail or an endpoint head

☞  $A \rightsquigarrow B$ :  $A \rightsquigarrow B$  or  $A \leftrightarrow B$

☞  $A \rightsquigarrow B$ :  $A \rightsquigarrow B$  or  $A \leftrightarrow B$  or  $A \leftarrow B$

**Wildcard** \* = concatenation of arcs (► Imagine '\*' as zero, one or more colliders)

☞ **Confounding path**  $A \leftrightarrow * \leftrightarrow B$ :  $A \leftrightarrow B$  or  $A \leftrightarrow \circ \leftrightarrow B$  or  $A \leftrightarrow \circ \leftrightarrow \dots \leftrightarrow B$

☞  $A \rightsquigarrow * \leftrightarrow B$ : path of any shape between  $A$  and  $B$

## Notation: refined m-connection

$A \langle \text{shape} \rangle B \mid C \iff \exists$  a path of  $\langle \text{shape} \rangle$  between  $A$  and  $B$  that is  $m$ -connected given  $C$ ,

$A \langle \cancel{\text{shape}} \rangle B \mid C \iff \nexists$  a path of  $\langle \text{shape} \rangle$  between  $A$  and  $B$  that is  $m$ -connected given  $C$ .

(► When  $C = \emptyset$ , ' $A \langle \text{shape} \rangle B \mid \emptyset$ ' is not be shortened to ' $A \langle \text{shape} \rangle B$ '.)

👉 **m-connection and m-separation** (► ' $\rightsquigarrow * \leftarrow$ ' is a path of any shape)

$A \rightsquigarrow * \leftarrow B \mid C \iff A \not\perp_m B \mid C$  and  $A \rightsquigarrow \cancel{*} \leftarrow B \mid C \iff A \perp_m B \mid C$ .

### Refined m-connection

$$A \left\{ \begin{array}{c} \rightsquigarrow \\ \rightsquigarrow \\ \rightsquigarrow * \leftarrow \end{array} \right\} B \mid C$$

' $\leftarrow$ ': **confounding arc** (► ' $\leftarrow$ ' / ' $\leftrightarrow$ ' in our procedure)

' $\rightsquigarrow * \leftarrow$ ': **confounding path**

## Back-door criterion, reformulated

**Back-door criterion, reformulated** For any  $S \subset V \setminus \text{de}(X)$ ,

$$S \text{ satisfies the back-door criterion} \iff X \leftrightarrow * \leftrightarrow Y \mid S.$$

(► Suppose  $X \rightsquigarrow Y$  in  $\mathcal{G}$ .)

👉 This is complicated by **collider bias**

$$A \leftrightarrow * \leftrightarrow B \mid C \not\Rightarrow A \leftrightarrow * \leftrightarrow B \mid C', \quad C \subset C'.$$

👉 However, **confounding arc** is free of such issue

$$A \leftrightarrow B \mid C \implies A \leftrightarrow B \mid C', \quad C \subset C'.$$

As our notation suggests, any m-connected confounding path ' $\leftrightarrow * \leftrightarrow$ ' is one or more m-connected confounding arcs ' $\leftrightarrow$ '.

👉 **Main idea**: block all confounding paths by **blocking confounding arcs**, one at a time.

## Primary adjustment set

Set  $C$  is a **primary adjustment set** for  $A, B$  relative to  $S$  if  $C \cap (\text{de}(A) \cup \text{de}(B)) = \emptyset$  and

$$A \leftrightarrow B \mid S, C.$$

When  $S = \emptyset$ ,  $C$  is simply called a primary adjustment set for  $A, B$ .

(►  $S$  is a posited adjustment set)

👉 To block  $X \leftrightarrow * \leftrightarrow Y \mid S$ , each time, we find a primary adjustment set for a constituent confounding arc and add it to  $S$ .

►  $\{X, Y\} \cup S$  is growing — how do we keep track of confounding relations?

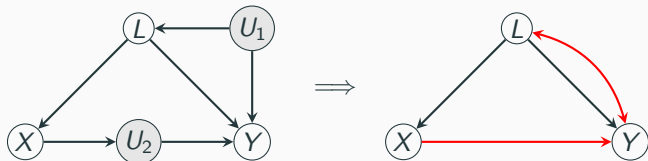
Questions?

# Latent projection

☞ Causal model represented by ADMG is closed under **latent projection**.

For  $V = \tilde{V} \cup U$  and  $\tilde{V} \cap U = \emptyset$ , let  $\mathcal{G}(\tilde{V})$  be the latent projection of  $\mathcal{G}$  onto the margin  $\tilde{V}$ .

$$A \left\{ \begin{array}{l} \text{via } U \\ \rightsquigarrow \\ \text{via } U \\ \leftarrow \\ \text{via } U \\ \leftrightarrow \end{array} \right\} B[\mathcal{G}] \iff A \left\{ \begin{array}{l} \rightarrow \\ \leftarrow \\ \leftrightarrow \end{array} \right\} B[\mathcal{G}(\tilde{V})].$$





# Preservation of refined m-connection

**m-connection is preserved by latent projection**

$$A \overset{\rightsquigarrow}{\longleftrightarrow} * \overset{\rightsquigarrow}{\longleftarrow} B \mid C [\mathcal{G}(\tilde{V})] \iff A \overset{\rightsquigarrow}{\longleftrightarrow} * \overset{\rightsquigarrow}{\longleftarrow} B \mid C [\mathcal{G}], \quad \{A, B\} \cup C \subseteq \tilde{V} \subseteq V.$$

**Theorem** For any ADMG  $\mathcal{G}$  over vertex set  $V$  and  $\{A, B\} \cup C \subseteq \tilde{V} \supseteq V$ ,

$$A \left\{ \begin{array}{c} \rightsquigarrow \\ \rightsquigarrow \\ \rightsquigarrow * \rightsquigarrow \end{array} \right\} B \mid C [\mathcal{G}] \iff A \left\{ \begin{array}{c} \rightsquigarrow \\ \rightsquigarrow \\ \rightsquigarrow * \rightsquigarrow \end{array} \right\} B \mid C [\mathcal{G}(\tilde{V})].$$

(► Relations induced by  $\rightsquigarrow$  and  $\rightsquigarrow \neq \rightsquigarrow$  are weaker than semi-graphoids.)

**Corollary** For  $A, B \notin C$ ,

$$A \left\{ \begin{array}{c} \rightsquigarrow \\ \rightsquigarrow \\ \rightsquigarrow * \rightsquigarrow \end{array} \right\} B \mid C [\mathcal{G}] \iff A \left\{ \begin{array}{c} \rightarrow \\ \leftrightarrow \\ \leftrightarrow * \leftrightarrow \end{array} \right\} B [\mathcal{G}(\{A, B\} \cup C)].$$

# Representation and procedure

## District criterion

$$\underbrace{S \text{ is a sufficient adjustment set}}_{\text{block all back-door paths in } \mathcal{G}} \iff \underbrace{X \leftrightarrow \not{*} \leftrightarrow Y \mid S [\mathcal{G}(\{X, Y\} \cup S)]}_{\text{connectivity by confounding arcs}}.$$

☞ This leads to an explicit **representation** and a **procedure**.

- For a posited  $S$ , for every  $u, v \in \bar{S} := \{X, Y\} \cup S$ , maintain **edges** as **m-connected confounding arcs** (► directed edges are irrelevant):

$$u \longleftrightarrow v \iff u \leftrightarrow v \mid \bar{S} \setminus \{u, v\}.$$

- An edge is initially drawn as a potential confounding arc ' $\leftarrow - \rightarrow$ '.  
If there exists a **primary adjustment set** to control it, the edge is removed;  
Otherwise, the edge becomes an uncontrollable confounding arc ' $\longleftrightarrow$ '.
- Primary adjustment set can be constructed from local knowledge about **common causes** and **mediators**.

1: $\mathcal{R} = \{\}$	▷ Set of sufficient adjustment sets
2: $\mathcal{Q} = \text{PRIORITYQUEUE}((\emptyset, \emptyset, \emptyset))$	▷ Initial graph has a possible edge $X \leftarrow - \rightarrow Y$
3: <b>while</b> $\mathcal{Q} \neq \emptyset$ <b>do</b>	
4: $(S, \mathcal{B}_y, \mathcal{B}_n) = \text{POP}(\mathcal{Q})$	▷ Pop a graph with smallest min-cut( $X, Y$ )
5: $\bar{S} = S \cup \{X, Y\}$	
6: <b>if</b> $X \leftrightarrow * \leftrightarrow Y$ by edges in $\mathcal{B}_y$ <b>then</b>	▷ Fails the district criterion
7: <b>continue</b>	
8: <b>else if</b> $X \leftrightarrow * \leftrightarrow Y$ by edges in $(\bar{S} \times \bar{S}) \setminus \mathcal{B}_n$ <b>then</b>	▷ Satisfies the district criterion
9: $\mathcal{R} = \mathcal{R} \cup \{S\}$	
10: <b>continue</b>	
11: <b>end if</b>	
12: $(A, B) = \pi = \text{SELECTEDGE}(X, Y, S, \mathcal{B}_y, \mathcal{B}_n)$	▷ $\pi$ is selected from $(\bar{S} \times \bar{S}) \setminus (\mathcal{B}_y \cup \mathcal{B}_n)$
13: $\mathcal{L} = \text{FINDPRIMARY}((A, B); S \setminus \{A, B\})$	
14: <b>if</b> $\emptyset \in \mathcal{L}$ <b>then</b>	
15: $\text{PUSH}(\mathcal{Q}, (S, \mathcal{B}_y, \mathcal{B}_n \cup \{\pi\}))$	▷ $\pi$ need not be expanded
16: <b>else</b>	
17: <b>for</b> $C \in \mathcal{L}$ <b>do</b>	
18: $\text{PUSH}(\mathcal{Q}, (S \cup C, \mathcal{B}_y, \mathcal{B}_n \cup \{\pi\}))$	▷ Expand $\pi$ by each primary adjustment set
19: <b>end for</b>	
20: $\text{PUSH}(\mathcal{Q}, (S, \mathcal{B}_y \cup \{\pi\}, \mathcal{B}_n))$	▷ Not to expand $\pi$
21: <b>end if</b>	
22: <b>end while</b>	
23: <b>return</b> $\mathcal{R}$	

## Soundness and completeness

☞ If running the procedure **exhaustively** (▶ with a stack or a queue) with the input answered by an **oracle** subroutine  $\text{FINDPRIMARY}((A, B); S')$ , the following can be shown.

**Theorem** Consider any two vertices  $X, Y$  in an ADMG  $\mathcal{G}$  such that  $X \rightsquigarrow Y$ .

1. **Soundness (primary  $\Rightarrow$  sufficiency)**: Suppose every  $C \in \text{FINDPRIMARY}((A, B); S')$  is a primary adjustment set for  $A, B$  given  $S'$  in  $\mathcal{G}$ . Then every element in the output of  $\text{CONFOUNDERSELECT}(X, Y)$  is a sufficient adjustment set for  $(X, Y)$ .
2. **Completeness (all minimal primary  $\Rightarrow$  all minimal sufficiency)**: Suppose further that  $\text{FINDPRIMARY}((A, B); S')$  contains all minimal primary adjustment sets for  $(A, B)$  given  $S'$  in  $\mathcal{G}$ . Then the output of  $\text{CONFOUNDERSELECT}(X, Y)$  contains all minimal sufficient adjustment sets for  $(X, Y)$ .

▶ 'Minimal': no proper subset is also a primary/sufficient adjustment set.

- Confounding **path** ' $\leftrightarrow * \leftrightarrow$ ' provides a structural definition of **confounding**.
  - ▶ Confounding arcs ' $\leftrightarrow$ ' are the building blocks.
  - ▶ Related: definition of a confounder (VanderWeele and Shpitser, 2013)
- **Economical queries**: causal relations between variables in  $S$  are never elicited.
- The user answers questions about **common causes** and **mediators**. Need not be aware of **collider bias**.
  - ▶ Compared to undirected graphs, colliders can be an extra complication to beginners in graphical models.
  - ▶ This was an issue of debate in the literature (Shrier, 2008; Rubin, 2009; Pearl, 2009; Sjölander, 2009).
- Moving away from the “known  $\mathcal{G}$ ” stance: typically, **identification** only relies on certain **partial knowledge** the underlying causal graph.
  - ▶ How to **represent** such knowledge? ▶ How to elicit it?
  - ▶ More development is needed on how to use causal graphs to **design** a study.











**Give it a try!**

<https://ricguo.shinyapps.io/InteractiveConfSel/>






arxiv: 2309.06053


## References

-  Guo, F. Richard, Anton Rask Lundborg, and Qingyuan Zhao (2022). “**Confounder Selection: Objectives and Approaches**”. In: *arXiv preprint arXiv:2208.13871*.
-  Pearl, Judea (1993). “**Comment: graphical models, causality and intervention**”. In: *Statistical Science* 8.3, pp. 266–269.
-  Shpitser, Ilya, Tyler J. VanderWeele, and James M. Robins (2010). “**On the validity of covariate adjustment for estimating causal effects**”. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. UAI'10. Arlington, Virginia, USA: AUAI Press, pp. 527–536.

-  Shrier, Ian and Robert W Platt (2008). **“Reducing bias through directed acyclic graphs”**. In: *BMC Medical Research Methodology* 8, pp. 1–15.
-  Haber, Noah A et al. (2022). **“DAG With Omitted Objects Displayed (DAGWOOD): a framework for revealing causal assumptions in DAGs”**. In: *Annals of Epidemiology* 68, pp. 64–71.
-  VanderWeele, Tyler J. and Ilya Shpitser (2011). **“A new criterion for confounder selection”**. In: *Biometrics* 67.4, pp. 1406–1413.
-  VanderWeele, Tyler J. (2019). **“Principles of confounder selection”**. In: *European Journal of Epidemiology* 34.3, pp. 211–219.
-  Richardson, Thomas S. (2003). **“Markov Properties for Acyclic Directed Mixed Graphs”**. In: *Scandinavian Journal of Statistics* 30.1, pp. 145–157.



-  VanderWeele, Tyler J. and Ilya Shpitser (2013). **“On the Definition of a Confounder”**. In: *The Annals of Statistics* 41.1, pp. 196–220.
-  Shrier, Ian (2008). **“Letter to the Editor”**. In: *Statistics in Medicine* 27.14, pp. 2740–2741.
-  Rubin, Donald B. (2009). **“Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?”** In: *Statistics in Medicine* 28.9, pp. 1420–1423.
-  Pearl, Judea (2009). **“Remarks on the method of propensity score”**. In: *Statistics in Medicine* 28.9, pp. 1415–1416.
-  Sjölander, Arvid (2009). **“Propensity scores and M-structures”**. In: *Statistics in Medicine* 28.9, pp. 1416–1420.

-  Sadeghi, Kayvan and Steffen Lauritzen (2014). “**Markov properties for mixed graphs**”. In: *Bernoulli* 20.2, pp. 676–696.

# Graphoid-like properties

► Let  $\mathcal{J}$  be a collection of tuples  $\langle A, B \mid C \rangle$  for  $A, B, C$  that are disjoint subsets of a ground set  $V$ . For disjoint  $A, B, C, D \subset V$ , consider the following properties:

- (i) triviality:  $\langle A, \emptyset \mid C \rangle$  for every disjoint  $A, C \subset V$ ;
- (ii) symmetry:  $\langle A, B \mid C \rangle \implies \langle B, A \mid C \rangle$ ;
- (iii) decomposition:  $\langle A, B \cup C \mid D \rangle \implies \langle A, B \mid D \rangle$  and  $\langle A, C \mid D \rangle$ ;
- (iv) weak union:  $\langle A, B \cup C \mid D \rangle \implies \langle A, B \mid C \cup D \rangle$ ;
- (v) contraction:  $\langle A, C \mid D \rangle$  and  $\langle A, B \mid C \cup D \rangle \implies \langle A, B \cup C \mid D \rangle$ ;
- (vi) intersection:  $\langle A, B \mid C \cup D \rangle$  and  $\langle A, C \mid B \cup D \rangle \implies \langle A, B \cup C \mid D \rangle$ ;
- (vii) composition:  $\langle A, B \mid D \rangle$  and  $\langle A, C \mid D \rangle \implies \langle A, B \cup C \mid D \rangle$ .

We say  $\mathcal{J}$  is a *semi-graphoid* over  $V$ , if it satisfies (i)–(v); further, we say  $\mathcal{J}$  is a *graphoid* over  $V$ , if it satisfies (i)–(vi), and finally, a *compositional graphoid* over  $V$ , if it satisfies (i)–(vii).

► m-separation  $\mathcal{J}_{\not\leftrightarrow} \neq \not\leftrightarrow$  is a compositional graphoid (Sadeghi and Lauritzen, 2014).

## Relations

$\mathcal{J}_{\not\leftrightarrow} := \{ \langle A, B \mid C \rangle : A \not\leftrightarrow B \mid C [\mathcal{G}] \}$  and  $\mathcal{J}_{\not\leftrightarrow} \neq \not\leftrightarrow := \{ \langle A, B \mid C \rangle : A \not\leftrightarrow \neq \not\leftrightarrow B \mid C [\mathcal{G}] \}$   
only satisfy properties (i)–(iv) and (vii).