Meeting notes 4 21 2023

Everyone was there except Dennis, who was off doing persuasive percussion.

I. Amber has the 'keep the best' for downstream running, though of course downstream-only means that 'keep the best' may actually be keeping something that's not quite right. That's part of why it's interesting.

All that remains in set-up for the 'comparison of heuristics' is implementing 'keep the best' for hybridization. It would be great if Dennis and Amber could work on that next.

II. Zhongming is taking another direction from that presented last week. He is still looking at matrix representations, but realized that many of the dynamics he was investigating demanded networks that looped.

III. Most of the time was taken up with Patrick giving a book report on Judea Pearl on the do-operator, back-door adjustment, front-door criterion, and some thoughts on how this might or might not interface with what we're doing.

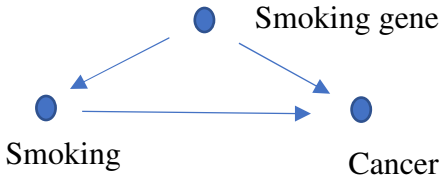
Here are the notes from that:

Pearl on do-operator, back-door adjustment, and front-door criterion

A. Background causal graph

We hypothesize a causal graph between variables. What we want to know is the causal effect of one variable on another—which may not be just 1 and 0, but the extent of the causal effect of one variable on another.

The causal graph shows hypothetical causal connections between variables. Pearl's example from debates over whether smoking causes lung cancer:



Everyone agreed on an observational correlation between smoking and cancer. Some people proposed that smoking causes cancer. R. A. Fisher proposed an alternative explanation of the data: that there was a 'confounder,' a gene that both inclined people to smoke and independently raised the probability of getting cancer.

Note that even here there are some assumptions built into the causal graph that we are working with: nobody proposed that having cancer caused you to have a smoking gene, for example.

The task: to find out what causal connections hold, and to what extent one thing causes another,if possible from observational (non-interventionist) data alone.

B. The do-operator

What we want to know in knowing whether X causes Y (and to what extent) is not just corelated values, but how the values of Y vary when we 'wiggle' X alone.

P(Y|X) is the probability of Y when X is observed. That isn't necessarily causal, as the example above would show if the smoking gene were the real cause.

What we want to know causally is P(Y|do (X)), where 'do' is the do-operator. We say that X causes Y if P(Y|do(X)) > P(Y).

"The do-operator erases all the arrows that come into X, and in this way it prevents any information about X from flowing in the non-causal direction. Randomization has does the same effect. So does statistical adjustment, if we pick the right variables to adjust" - *The Book of Why*, p. 157

C. The back door adjustment

So given the assumptions represented in a causal graph, how do we know whether X really causes Y? One way is to intervene on X.

How can we tell whether X really causes Y without intervention, from observational data alone?

We do that by simulating the do-operator, and we can do that by back-door adjustment, blocking all the non-causal 'backdoor' paths between X and Y.

Two parts to this: How do we know the non-causal paths?
How do we block them?

How do we know the non-causal paths? Start with all paths with arrows into X. Trace those back to Y. Those are all the paths we have to block.

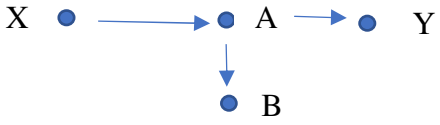
How do we block them? If in a path we have a chain A → B → C or A ← B → C (chain or fork), we can block the path by controlling for B.

If in a path we have a collider A → B ← C, that path is already blocked.

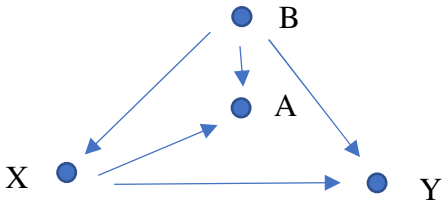
Blocking back-door paths also amounts to 'deconfounding' the variables in question, and is also referred to as 'd-separation.'

> An additional wrinkle: controlling for descendants of a variable is like partially controlling for the variable itself. You have to watch out for that.

Examples from Pearl:



Does X cause Y? There are no backdoor routes from X, so you don't have to control for anything.

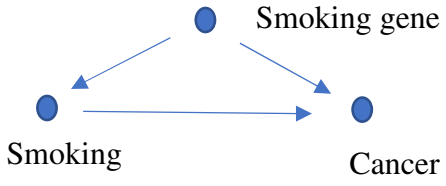


Does X cause Y? There is one backdoor route from X to Y, through B. That path is of the form A ← B → C, so we can block it by controlling for B.

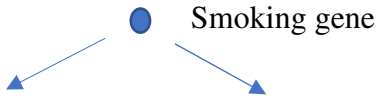
That's the idea. And how do you control for a variable? For discrete variables (like gender, perhaps) you find the effect for each value and average them by percentages. More details on more complicated adjustments are in Pearl, 221 Even more gory details in Pearl, Glymour, and Jewell, *Causal Inference in Statistics: A Primer.*

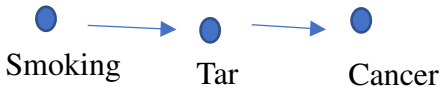
D. Front door criterion

This won't do everything. In particular, this demands controlling for variables, which must be observed. If unobservable, won't work. That's the problem with the smoking case—Fisher's postulated 'smoking gene' was an unobservable. Since we don't have values for it, we can't control for it, and so can't block the possible 'back door.'



In that case, one can sometimes use a 'front door' route. Pearl laments that this hasn't been used more. In the smoking case, suppose we also have data on tar levels in the lung, and the assumptions on the table can be represented in this causal graph. Nobody proposed that the smoking gene directly causes tar, for example.





Here we first establish a causal connection between smoking and tar. A backdoor from smoking to tar is blocked because it would have to go through the 'collider' at cancer, which blocks it.

Then we establish a causal connection between tar and cancer, with a backdoor from tar similarly blocked.

We calculate the smoking to cancer causality from those.

But note that without that convenient extra variable, there is no way of eliminating the 'smoking gene' confounder from observational data alone.

E. How does this connect to our stuff?

Ways that it doesn't connect:

- This assumes a causal graph, which embeds certain assumptions. That graph can tell you where to intervene to establish causality, and can also tell you how to simulate intervention from observational data alone.
    - We are looking for building a causal graph from the data to begin with, or at least altering an existing one from the data. We're interested in changing graph. We envisage a scientific theory 'flailing' in order to find the world. This sort of assumes that scientific theory already in place. The Pearl approach never changes the assumed graph (though perhaps it could, if it eliminated the smoking gene, for example).
- This can deal with extent of causal influence, rather than just 1 or 0.
    - Ours can't, but we've thought of that on the horizon.
- This works from static 'spreadsheet' data, not timed data.
    - We are exploring both 'downstream' data, like this, *and* timed data, looking for potential differences.
        But note that timed data doesn't solve everything: a 'confounder' might have a delay. Fisher might propose that a gene causes a desire to smoke in young people, for example, and independently an increase in probability of cancer in older people.

Ways that it might connect:

●○ We have established causality essentially by intervention – activating nodes. Perhaps these techniques would supplement what we could learn from some interventionist data without having all interventionist data.

●○ Related: We've been interested in how much you lose if you have only partial interventionist data. This might be relevant to that, if sometimes you don't lose by losing interventionist data, because it can be reconstructed using these techniques.

IV. For next time, whenever that is:

I hope Amber and Dennis will continue work on 'keep the best' (both timed and downstream) for the hybridization heuristic.

Given the end of the semester and the beginning of summer, we'll have to figure out people's schedules in order to see how to meet to continue the work. We know that:

Amber will be away from April 28th to June 11th

Zhongming will be away from April 27th to June 20th.

We will find out about Dennis nd Sophia's plans and see what works.