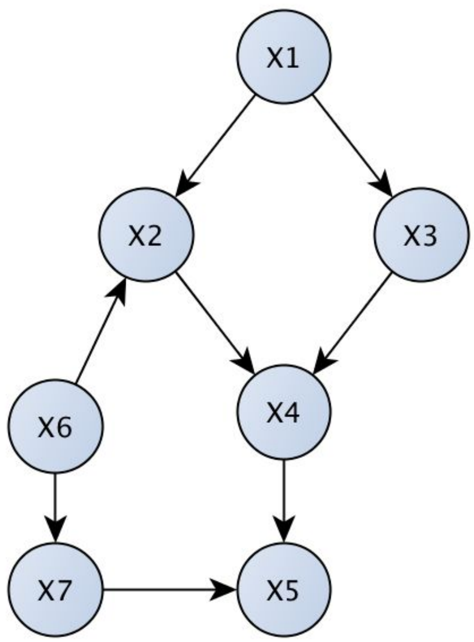
## **Homework 2**: Causal Graphs, d-separation, and the back-door criterion

(1) You have a joint distribution on three variables, $P(X_1, X_2, X_3)$. You've measured all dependencies in this distribution, and have only found one conditional independence relationship, $X_1 \perp X_2 \mid X_3$. Draw all possible causal graphs (assume they're acyclic) that can represent this distribution.

This implies that $X_3$ separates $X_1$ and $X_2$ so they're not directly connected. We don't have $X_1$ and $X_2$ being statistically independent, so what we're left with is the 3-node, 2-edge graphs $X_1 \leftarrow X_3 \rightarrow X_2$, $X_1 \rightarrow X_3 \rightarrow X_2$, and $X_1 \leftarrow X_3 \leftarrow X_2$.

(2) In the following graph, give a set Z that d-separates




(a) $X_1$ and $X_7$

There are paths

i) X1 → X2 ← X6 → X7

ii) X1 → X2 → X4 → X5 ← X7

iii) X1 → X3 → X4 → X5 ← X7

iv) X1 → X3 → X4 ← X2 ← X6 → X7

It's tricky since X2 is a collider -- you'd like to just cut off X1 by conditioning on its children, but that doesn't work! Instead, lets look at the paths one by one.

The first is blocked by X6 and no others, or by not conditioning on X2. We'll have to include X6, or avoid conditioning on X2.

The second is blocked by X2 or X4 (or both), or by not conditioning on X5. We'll have to include at least one of these, or exclude X5.

The third is blocked by X3 or X4, or excluding X5. We'll have to include at least one of these, or exclude X5..

The last is blocked by X3, X2, or X6, or by not conditioning on X4. We'll have to include at least one of these, or exclude X4.

Let's pick Z={X6, X2, X4}. We still have to check it to make sure we didn't unblock any colliders. X2 and X4 are both colliders! Did we rely on them to block any paths? We didn't! We picked variables from the set of things to condition on to block each path, so we're okay!

(b) $X_2$ and $X_5$

The paths are
(i) X2 → X4 → X5
(ii) X2 ← X6 → X7 → X5
(iii) X2 ← X1 → X3 → X4 → X5

The first is blocked by X4. The second is blocked by X6 or X7. the third is blocked by X1, X3, or X4. We can pick Z={X4, X6}, and then check each path. We didn't rely on any colliders, so we should be good!

(c) $X_1$ and $X_5$

(i) X1 → X3 → X4 → X5
(ii) X1 → X2 → X4 → X5
(iv) X1 → X2 ← X6 → X7 → X5

X4 or X3 blocks the first path. X2 or X4 block the second. X6, X7 block the third, or avoiding conditioning on X2. At first glance, it looks like we can pick Z={X4}, but notice that X4 is a descendant of X2. This unblocks the collider at X2! Z={X4, X6} would do the trick.

(3) In the graph from (2), give a Z that satisfies the back-door criterion for the causal effect of

(a) $X_1$ on $X_7$

There are no back-door paths, because there are no arrows into X1. Z={} satisfies the back-door criterion!

(b) $X_2$ on $X_5$

We enumerated the paths in the last problem:

(i) X2 → X4 → X5

(ii) X2 ← X6 → X7 → X5

(iii) X2 ← X1 → X3 → X4 → X5

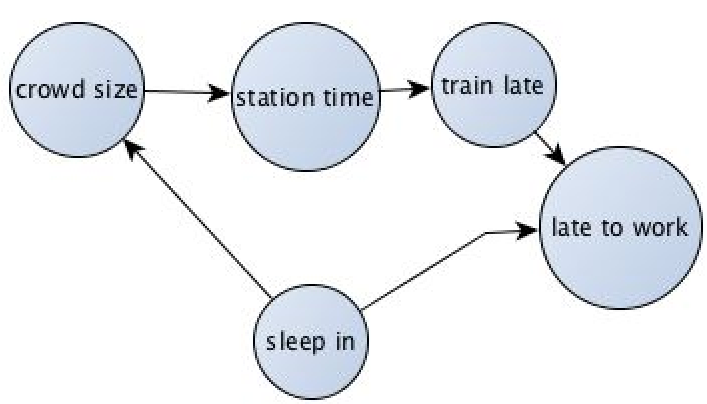
Of these, (ii) and (iii) have arrows pointing into X2. We need to block these paths, while avoiding breaking the back-door criterion (i.e. we shouldn't condition on descendants of the causal state). The descendants of X2 are X4 and X5, so we're not allowed to control for these. Then, to block paths (ii) and (iii), we can control for X6 and X3, so Z={X6, X3} satisfies the back-door criterion for the effect of X2 on X5.

(4) Explain qualitatively how the sets from (2) and (3) are different. What is important about the paths left unblocked in (3) that were blocked in (2)?

In question (2) we block all paths between the variables, whereas in (3) we only block certain paths: ones with arrows pointing into the causal state. If these paths drive statistical association between the causal state and the outcome, then it must be spurious association, because causation is going in the wrong direction. The remaining paths, if they result in statistical association between the causal state and the outcome, must drive "causal" statistical association, that is, statistical association resulting from directed paths.

(5) Draw a causal graph for each of the following systems:

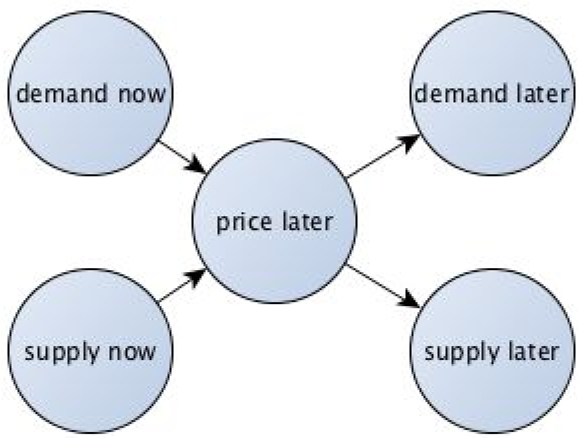
(a) Large crowds cause trains to stay in the station longer. Staying in the station longer will make trains are late. If trains are late, you will be late for work. When you sleep in, you are late for work, but when you sleep in, the crowds on the train are smaller.



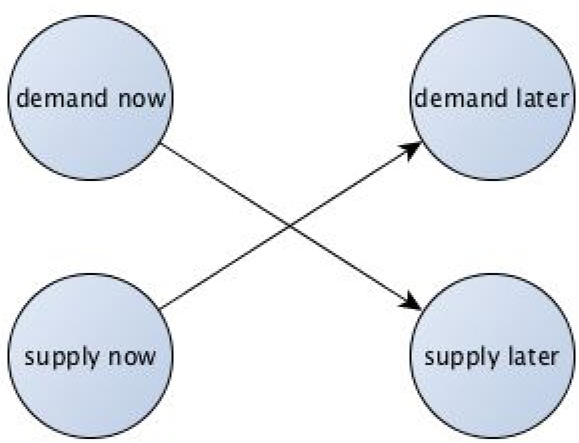
This is the graph, but notice that the causal states aren't finely articulated. Do we mean the average time a train is in the station? Do we mean the average station time for the train I end up taking to work? If we really wanted to study this problem, we'd have to define these much more carefully.

(b) When demand for a product is high (i.e. many people want to buy it), the price goes up because vendors know they can charge more. When prices increase, demand decreases because fewer people are willing to buy expensive products. (Hint: Supply now will drive demand later. Demand now will drive supply later.)

This one is tricky. There are a lot of ways you can draw it, like




Or you might suppress the price part of it, and write




The important aspect is realizing that instead of creating a cyclic graph where supply effects demand, which effects supply, you can look at these variables over time in a DAG.