252D - My Part

1. Doing the dag

   Before analyzing our directed acyclic graph we thought it appropriate to describe the thought process that led us to it. We started out with a complete graph with edges between all nodes, where each node is a separate variable. The variables in our dataset were:

   - the employee satisfaction level (satisfaction)
   - the score on the employee's last (most recent) evaluation (last_evaluation or evaluation)
   - the number of projects an employee has completed while at the company (#projects)
   - the average monthly hours (avg_monthly_hours or hours)
   - the time an employee has spent at the company (time_spend_company)
   - whether the employee had a work accident or not (work_accident or accident)
   - whether the employee was promoted in the last 5 years (promotion)
   - the employee's department (department)
   - the employee's salary (salary)
   - whether the employee left (left).

   In the original graph we could identify 4 exclusion restrictions: Between department and promotion, between accident and promotion, between salary and accident and between Department and last evaluation. Except for these four edges our initial graph would still be fully connected. However, we soon stumbled on other difficulties. For example, we found reason for an edge between avg_monthly_hours and work_accident (working more hours could potentially increase the risk of a work

accident) but also the other way around (after an accident people might work less). We therefore decided that we do not want to assume directionality between these two nodes, so we decided to combine them. It is important to note that we did not combine them because we knew that the interaction is bidirectional, we did so because we could not exclude the possibility that it is. We stumbled upon the same problem when looking at the nodes #projects and average_monthly_hours. Both could potentially influence each other. Because we already combined work_accident and avg_monthly_hours, we now had to combine all three of those variables. Because of the way #projects is defined (number of projects completed while in company) it might be influenced by time_spend_in_company (someone who has been at the company for a longer time might have been able to complete more projects). However, the time spent in the company is probably influenced by whether someone had a work accident. Because work_accident and #projects have been combined to a node, there would now be an arrow between time_spend_in_company and the combined node, and between the combined node and time_spend_in_company. The resulting graph would therefore not be aciclyc anymore, and we had to combine the already combined node and time_spend_in_company. The same logic applied for the variable promotion. Additionally, we felt that we should also combine the varibales last_evaluation and satisfaction, again because we could not say for sure that the edge connecting the two would be one directional. So, our "true" causal graph consisted of a combined node promotion_projects_accidents_timeInCompany_avgHours (from now on referred to as PPATH), of the evaluation_satisfaction node, of the department node, as well as the salary and left node. We furthermore decided that the department might influence for example the avg_monthly_hours, but not the other way around and that the department might influence the employee's satisfaction (CHECK THIS WITH OTHERS! DIRECT INFLUENCE?!?) the salary, and whether someone leaves. Whether someone leaves is in turn also affected by salary,

the department and the combined node PPATH. We also decided that at least in the initial, "true" causal graph no independence assumptions were warranted.
We were aware of the fact that drawing a "W-A-Y" DAG might have saved us a lot of work and contain our graph as a special case, but decided that the exercise would proof its worth later in the roadmap.

2. The scientific question we want to answer is whether an increase in salary leads to a change in the average probability of leaving (check if everyone agrees with this phrasing). Specifically, because salary has 3 levels "low", "medium" and "high" we first concentrate on the effect an increase from "low" to "high" salary has. EDIT THIS IF WE DO MORE!

3. We will refer to "department" as "D", "evaluation_satisfaction" as "EA", "left" as "L", "salary" as "A" and "promotion_projects_accidents_time_hours" as before (PPATH)
The structural equations that follow from this are:

$$
\begin{aligned}
f_D &= f(U_{D,ES}, U_{D,PPATH}, U_{D,A}, U_{D,L}) \\
f_{PPATH} &= f(D, U_{D,PPATH}, U_{A,PPATH}, U_{ES,PPATH}, U_{L,PPATH}) \\
f_A &= f(PPATH, D, U_{A,D}, U_{A,PPATH}, U_{A,L}, U_{A,ES}) \\
f_{ES} &= f(PPATH, D, A, U_{A,ES}, U_{D,ES}, U_{PPATH,ES}, U_{L,ES}) \\
f_L &= f(PPATH, D, A, U_{D,L}, U_{A,L}, U_{PPATH,L}, U_{ES,L})
\end{aligned}
$$

4. Defining target causal parameter We define our target causal parameter $\Psi^F(P_{U,X})$ (CHECK THIS! DEFINE X AND U!)

$\Psi^F(P_{U,X}) = E_{U,X}(Y_{high} - Y_{low})$, where $Y_{salaryLevel}$ are counterfactual outcomes for salary level "salaryLevel".

Final Project Rough Draft

April 14, 2017

1 Specify your observed data.

The observed data consists of n i.i.d copies of the random variable $O = (W_1, W_2, A, Z, Y)$, where W_1 consists of the following:

- Whether the employee was promoted in the last five years (promotion_last_5years)
- Number of projects completed while at work (number_project)
- Whether the employee had a workplace accident (Work_accident)
- Number of years spent in the company (time_spend_company)
- Average monthly hours at workplace (average_monthly_hours)

W_2 is department in which the employee works for (sales), A is relative level of salary (salary), Z is evaluation of employee performance (last_evaluation) and level of satisfaction (satisfaction_level), and Y is whether the employee left the workplace or not (left).

There are several unblocked backdoor paths from outcome Y to exposure A. See Figure 1.

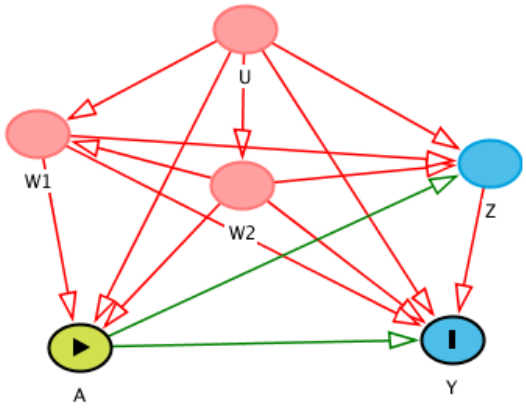
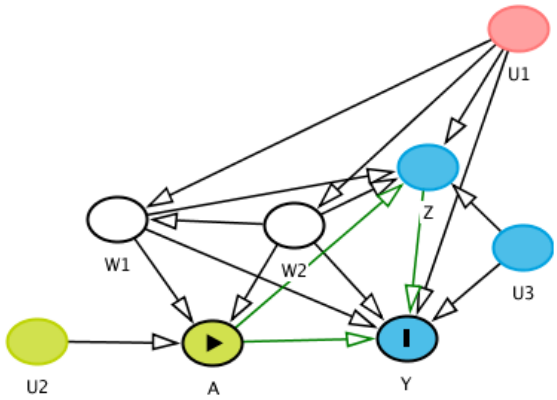


Figure 1: The unrestricted structural causal graph indicates that identifiability is not possible because the back-door criteria is not met.

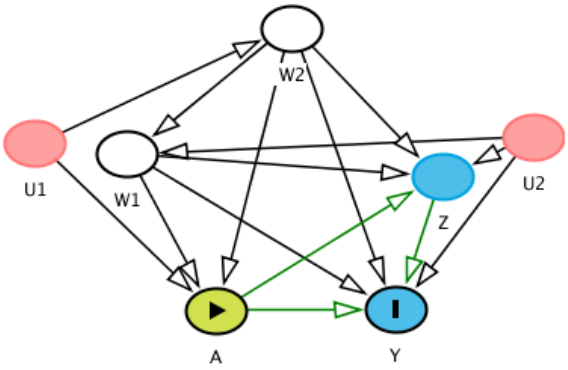
To fulfill the backdoor criteria, we would need to place independence assumptions on the distribution of unmeasured factors P_U. Specifically, there are four ways that the causal parameter can be identified as some parameter of our observed data distribution. Then, when we condition on W_1 and W_2, we would meet the back-door criteria.

Table 1: Four Possible Sets

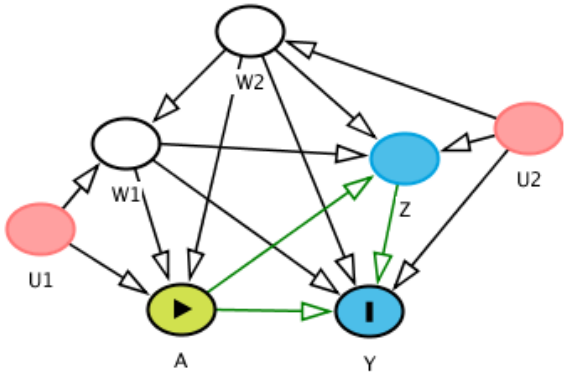
Set 1	Set 3
$U_{W_2} \perp U_Z$	$U_{W_1} \perp U_{W_2}$
$U_{W_2} \perp U_Y$	$U_{W_1} \perp U_Z$
$U_{W_1} \perp U_{W_2}$	$U_{W_1} \perp U_Y$
$U_A \perp U_Z$	$U_A \perp U_{W_2}$
$U_A \perp U_{W_1}$	$U_A \perp U_Z$
$U_A \perp U_Y$	$U_A \perp U_Z$
Set 2	**Set 4**
$U_{W_2} \perp U_Z$	$U_A \perp U_{W_2}$
$U_{W_2} \perp U_Y$	$U_A \perp U_Z$
$U_{W_1} \perp U_Z$	$U_A \perp U_Y$
$U_{W_1} \perp U_Y$	$U_A \perp U_{W_1}$
$U_A \perp U_Z$	
$U_A \perp U_Y$	



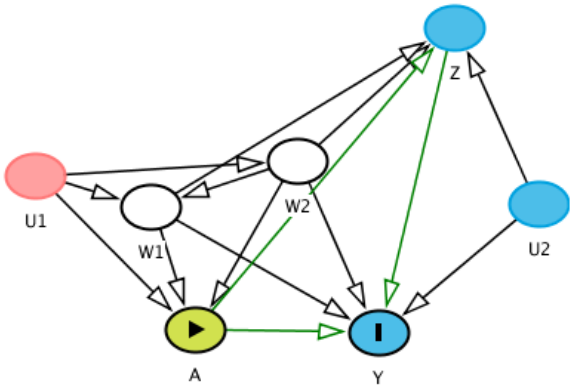
(a) Set 1



(b) Set 2



(c) Set 3



(d) Set 4

Figure 2: Possible sets of assumptions that meet the back-door criteria after conditioning on W_1 and W_2