1 Causal Influence as Controllability

This section expands section **A.4** from the Janzing et al 2013 paper. We provide an interpretation of the causal influence as the expected reduction in uncertainty from being allowed to choose the value of a variable – after other observed variables are chosen, but before the hidden noise variables. This helps provide an intution for the causal inference of an arrow, which will help give us an intuition for some of the surprising consequences of our generalization to the causal influence of a path.

Consider a variable Y with parents X and Z. It follows from Lemma 3 of the 2013 paper that $\mathcal{C}_{X\to Y}$ can be written

$$\mathbb{E}_{X,Y,Z} \log \frac{P(y|x,z)}{\sum_{x'} P(x')P(y|x',z)} \tag{1}$$

$P(y|x,z)$ is, of course, the probability of getting $Y=y$ if X and Z are already chosen. The $\sum_{x'} P(x')P(y|x',z)$ can be interpreted as the probability of getting $Y=y$ if Z is already chosen but X is chosen randomly. We can hence see $\log \frac{P(y|x,z)}{\sum_{x'} P(x')P(y|x',z)}$ as the reduction in uncertainty of the event $Y=y$ from choosing $X=x$.

Consider the example given by the structural equations

$$\begin{aligned} z &= \text{rand}(\{0,1\}) && (2)\\ x &= z && (3)\\ y &= x \oplus z && (4) \end{aligned}$$

where $\oplus$ is the XOR operation. There are two assignments of nonzero probability: $X=Y=Z=0$, and $X=Z=1, Y=0$. Consider the $X=Z=1$ case; the other is symmetric. Here, $\sum_{x'} P(X=x')P(Y=1|Z=1,X=x') = \frac{1}{2}$, indicating that, given we have picked $Z=1$, there is a $\frac{1}{2}$ probability of obtaining $Y=0$ if we randomly choose X. The $P(Y=0|X=1,Z=1)=1$ term means there is probability 1 of obtaining $Y=0$ if we further allow ourselves to pick $X=1$. Hence, if $Z=1$, then choosing $X=1$ gives us 1 bit of uncertainty reduction in obtaining $Y=0$. Averaging with the other case gives us a causal influence of $\mathcal{C}_{X\to Y}=1$.

We now consider the hidden noise variables. Suppose Y is a deterministic function of X, Z, and U, but U is hidden. Then we can write $P(y|x,z)$ as $\sum_u P(u)P(y|x,z,u)$. We can write $\mathcal{C}_{X\to Y}$ as

$$\mathbb{E}_{X,Y,Z} \log \frac{\sum_u P(u)P(y|x,z,u)}{\sum_{x',u} P(x')P(u)P(y|x',z,u)} \tag{5}$$

The top term can now be interpreted as the uncertainty in obtaining $Y = y$ if u is chosen randomly, while the bottom term is the uncertainty if both X and U are chosen randomly. This is the sense in which we should see ourselves as picking X after the other observed variables, but before the hidden noise variables. In particular, this means that, in general, changing the causal model so that a hidden variable becomes observed can change the causal influence of other variables.

2 Indirect Causal Influence

2.1 Impact of Path Deletion

Similar to how the causal influence of an edge is defined using a notion of the impact of edge deletion, we will define the indirect causal influence of a path by creating a notion of path deletion.

The impact of an edge deletion is defined by the Kullback-Leibler divergence between a distribution and a distribution where that edge has been "cut," meaning its source has been replaced with an independent copy of the source variable. We proceed similarly in defining the impact of path deletion. To create the modified distribution, we replace a path with a new path where the source node X has been replaced with an independent copy X', and all intermediate nodes Z_i are replaced with the value that Z_i would have been had X had the value of X'.

This modified distribution is most easily defined using the twin network method for counterfactuals. Given a path $X \to Z_1 \to \cdots \to Z_k \to Y$, we create new nodes $X', Z'_1, \ldots, Z'_k$, with edges between each, and replace the edge $Z_k \to Y$ with an edge $Z'_k \to Y$. X' will be distributed IID to X. However, each Z'_k will be a deterministic function of the previous node in the path and the other parents of the original Z_k – including the hidden noise term for Z_k. To obtain a distribution in the original variables, we simply marginalize out the primed variables.

Consider the graph in Figure 1 for a probability distribution which factors as $P(X, Z_1, Z_2, Y) = P(X)P(Z_1|X)P(Z_2|X)P(Y|Z_1, Z_2)$. Figure 2 shows the dual-network graphs for deleting the sets of paths $\{Z_1 \to Y\}, \{X \to Z_1 \to Y\}$, and $\{X \to Z_1 \to Y, Z_1 \to Y\}$ respectively.

Marginalizing out the primed variables, the first and third graphs factor as $P(X, Z_1, Z_2, Y) = P(X)P(Z_1|X)P(Z_2|X)P(Y|Z_2)$, with $P(y|z_1, z_2) = \sum_{z'_1} P(z'_1)P(y|z'_1, z_2)$ and the other terms the same. However, for the second graph, the unobserved noise term U_1 introduces a back-door dependence on the original Z_1. Because we cannot condition on the unobserved term,

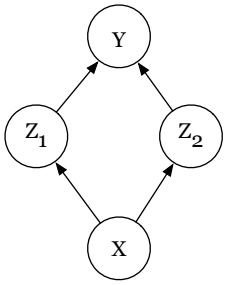
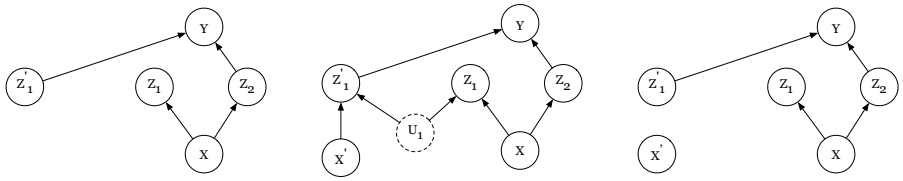


Figure 1



(a) (b) (c)

Figure 2

we must condition on the original Z_1. However, doing this introduces a further dependence on the original X, so we must condition on X as well. We thus must instead replace the $P(Y|Z_1, Z_2)$ term with $P(Y|z_1, z_2, x) = \sum_{z'_1, x,} P(x')P(z'_1|z_1, x, do(x'))P(y|z_1, z_2)$.

Terms like $P(z'_1|z_1, x, do(x'))$ are called counterfactuals. This term can be read "the probability that Z_1 would take the value z'_1 had X been x', given that the actual values of Z_1 and X are z_1 and x." It can be computed by first conditioning on Z_1 and X, using that to infer information about their ancestor variables (including the hidden noise terms), and then setting X to x' and computing forward. Another way to compute them is by constructing a twin network exactly like the one we derived it from, so that computing the counterfactual reduces to just conditioning.

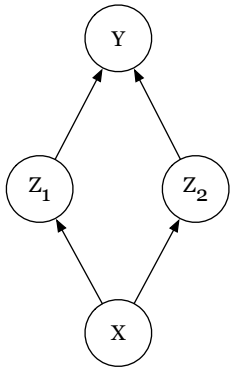


Figure 3

2.2 Paths: To Join or Not to Join?

Consider Figure 3. What is the total influence of the paths $X \to Z_1 \to Y$ and $X \to Z_2 \to Y$? We will show that it depends on whether we wish to consider them together or separately.

Figure 4a shows two possible dual networks for deleting the impact of these paths: one where both paths share a copy of X, and one where they each get a separate independent copy.

Suppose the graph in Figure 3 followed the following structural equations

$$X = \text{rand}(\{0,1\}) \tag{6}$$
$$Z_1 = X \tag{7}$$
$$Z_2 = X \tag{8}$$
$$Y = Z_1 \oplus Z_2 \tag{9}$$

(In this example, the hidden noise variables U_{Z_1} and U_{Z_2} are trivial.)

In this case, the influence of the set of paths $\{X \to Z_1 \to Y, X \to Z_2 \to Y\}$ is 1 bit because we can choose values flowing down each path so as to completely control Y. However, the causal influence of the subgraph with edges $\{X \to Z_1, X \to Z_2, Z_1 \to Y, Z_2 \to Y\}$ is 0 because we cannot change the value of Y by changing the value of X.

This phenomenon means that, when we give our general definition of indirect causal influence below, it is not enough to provide a set of paths.

Instead, we will need to provide a set of path sets, where each path sets is a set of path originating from a single node.

2.3 Consistency of Overlapping Paths

The previous section shed some light on situations where we have multiple copies of the source variable in a path. What if we have multiple copies of an intermediate variable?

Consider Figure 5. We wish to create a dual network for deleting the impact of the paths $\{X \to Z \to Y_1, X \to Z \to Y_2\}$.

Figure 6 shows two candidate twin networks for deleting the impact of these paths. Which one is preferable?

In some sense, it shouldn't matter, because Z is simply a deterministic function of U_Z and X: Z' and Z'' should always take on the same values. In this case, we could express this by merging them together into a single node. However, this is not an option in a more general case: we could not do this if there were an additional variable W such that W was a parent of Z' but W' was the corresponding parent of Z''. Yet, if W had no impact on Z, this would be the same example as in Figures 5 and 6.

It indeed does not matter whether we use the twin networks in Figure 6a or Figure 6b, as long as, for the former case, we take into account the dependence between Z' and Z'' through U, by e.g.: giving Z'' a term $P(Z'' = z''|Z' = z', Z = z, X = x, X' = x')$. But how can we express this once we have marginalized out Z' and X''? Because we expressed Z' as a counterfactual, this means we will need to condition on a counterfactual.

Condioning on counterfactuals is well-studied, but requires us to use different notation. We use the potential response notation of Judea Pearl, where the notation Y_x is used to denote the value of Y had X been x, so that, e.g.: $P(Y_x = y)$ is equivalent to $P(Y = y|do(X = x))$. We can now write the distribution of the resulting graph:

$$P(Y_1, Y_2|X, Z) = \sum_{x', z', z''} P(x')P(z'|z, x, do(x'))P(z''|z, x, Z_{x'} = z', do(x'))P(y_1|z')P(y_2|z'') \tag{10}$$

We can see immediately that the product $P(z'|z, x, do(x'))P(z''|z, x, Z_{x'} = z', do(x'))$ is 0 unless $z' = z''$. Hence, because we correctly modeled the dependence between Z' and Z'', it does not matter whether we use separate copies of Z in each path.

2.4 Definition

We now provide the general definition for the causal influence of paths. As we shall see, providing a definition that captures all the subtleties raised above will be quite cumbersome. The general approach is the same as used for defining the causal influence of an arrow:

1. Define a new probability distribution where the impact of paths has been deleted
2. Define the causal influence of those paths as the Kullback-Leibler divergence between that distribution and the original

We follow this general strategy for defining the new probability distribution:

1. Input a set of path-sets, where the paths in each path-set all originate from the same node. Paths originating from different variables can always be treated separately; however, in different situations we may want to treat paths from the same variable together or apart.
2. Transform the input into a set of path-sets where the paths in each path set all terminate at the same node . The input now resembles a set of trees, where the roots are the original nodes in the graph (final things being influenced in a path), except that the leaf nodes —the copies of variables which can be twiddled independently —are shared between trees. This is done because we cannot have two copies of a node if they are both meant to serve as a parent to the same node. While we may be able to introduce additional sharing between intermediate nodes, we can add conditioning as described above so that it doesn't matter. Despite the sharing, we will nonetheless refer to these graphs as "trees."
3. Write the joint probability distribution. When written in the form created in the previous step, there is enough information available to declaratively express the joint probability distribution, treating the source nodes, terminal nodes, and intermediate nodes of paths as the three distinct cases.

We will now begin the formal definition. We first define the transformation from lists of path-sets into tree sets. It will be helpful to consider an example when reading this definition: the list of path sets ($\{X \to Z_1 \to$

$Y, X \to Z_2 \to Y\}, \{X \to Y\}, \{X \to W \to Z_1 \to Y\})$ in any graph that has the corresponding edges will be transformed into a singleton set containing the tree shown in Figure 7

We use S to denote a list of path-sets, where S_i denoets the ith path-set. We can then define the corresponding tree set as

$$T = \{T_Y\} \tag{11}$$
$$T_Y = (V_Y, E_Y) \tag{12}$$

Before proceeding, we will need to define this auxilary function to give us the appropriate transformed version of a variable:

$$\text{trans} : \mathbb{N} \times \text{Path} \times \text{Var} \to \text{Var} \tag{13}$$
$$\text{trans}(i, p, X) = X \text{ if } X = \text{Dest}(p) \tag{14}$$
$$\text{trans}(i, p, X) = X^{(i)} \text{ if } X = \text{Source}(p) \tag{15}$$
$$\text{trans}(i, p, X) = X^{Y} \text{otherwise} \tag{16}$$

We use the shorthand trans$(i, p, X \to Z)$ to denote the edge $\text{trans}(i, p, X) \to \text{trans}(i, p, Z)$

Now we can define

$$E_Y = \{\text{trans}(i, p, X \to Z) | p \in S_i, \text{Dest}(p) = Y, X \to Z \in p\} \tag{17}$$
$$V_Y = \text{nodes}(E_Y) \tag{18}$$

We need a couple more ingredients before we can state the general form of the modified distribution. Fix an arbitrary variable ordering $X_1 < \cdots < X_n$, where $\{X_1, \ldots, X_n\}$ are the original variables of the graph. We then define V' as the newly-introduced variables.

$$V' = (\bigcup V_X) \setminus \{X_1, \ldots, X_n\} \tag{19}$$

The overall modified distribution takes the form:

$$P_S(X_1, \ldots, X_n) = \sum_{v'} PX * PY * PZ \tag{20}$$

The PX term represents the probability of the copies of the independent variables. It can be written simply

$$PX = \prod_{x^{(i)}} P(x^{(i)}) \tag{21}$$

where $x^{(i)}$ ranges over all variables of the form $X^{(i)}$ in V'.

The PY term gives the probability of the original variables in the new graph. Let $PA_Y^{T_Y}$ be the parents of Y in the tree T_Y, and $PA_Y^{\overline{T_Y}}$ be the parents of Y in the original graph, minus those variables with a corresponding variable in T_Y. Then we can write

$$PY = \prod_{y} P(y|pa_Y^{\overline{T_Y}}, pa_Y^{T_Y}) \tag{22}$$

where y ranges over the original variables in the graph.

Finally, the PZ term gives the probability of the intermediate variables in the paths. Here, for each modified Z in a tree, we infer the original value of its unknown noise variable by conditioning on the original Z and its parents, maintain consistency with the previously-seen modified versions of Z by conditioning on their respective counterfactuals, and counterfactually set its parents in the tree.

$$PZ = \prod_{z^Y} P(z^Y|z, pa_z, (Z_{pa_{Z^w}^{T_W}} = z^w)_{W<Y}, \mathrm{do}(pa_z^{T_Y})) \tag{23}$$

where z^Y ranges over all variables of the form Z^Y in V'.

Now, we can define the causal influence of S as simply $D(P||P_S)$.

2.5 Sensitivity to Functional Form

In the Janzing et al paper, we could ignore the need to specify deterministic functions and noise variables in causal models, and simply work in terms of conditional probabilities $P(x|z)$. We now show that, for the indirect causal influence, we will in general need to explicitly model the hidden noise terms: two causal models may have identical conditional probability tables, but different causal influences. The short explanation is that counterfactuals such as $P(z'|z, do(x))$ are sensitive to the exact form of the dependence on the hidden noise terms; the following example should make this more clear.

Example 1 (Noisy Copy A)**.** *Consider the graph in Figure 8. Suppose X is uniform 0 or 1, Y is a perfect copy of Z, and $X \to Z$ implements a "noisy copy" operation, so that $P(Z = 1|X = 1) = 0.75$ and $P(Z = 1|X = 0) = 0.25$. Figure 9 gives the corresponding twin-network for deleting the impact of the path $X \to Z \to Y$.*

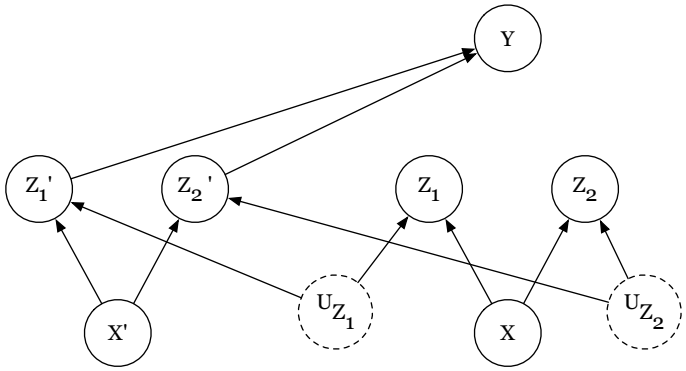
We now provide two models implementing this distribution. In model 1, $Z = X \oplus U_Z$, *where* U_Z *is a binary random variable. In model* 2, U *consists of two independent random binary variables* U_1, U_2, *and* $Z =$ *if* U_1 *then else* U_2.

The indirect causal inference of the path $X \to Y \to Z$ *is given by the KL divergence between the distribution of the twin network and the original distribution, namely* $\mathbb{E}_{X,Z,Y} \frac{P(y|z)}{\sum_{x',z'} P(x')P(z'|x,z,do(x'))P(y|z')}$. *In model* 1, *we can solve for* $U_Z = X \oplus Z$, *giving* $Z' = X' \oplus X \oplus Z$. *We can use this to evaluate* $P(z'|x, z, do(x')) = 1$ *iff* $z' = x \oplus z \oplus x'$, *resulting in a causal influence of* 1. *In model 2, however, if* $x = z$, *then, for* $z' = x' = \neg x$, *we only have* $P(Z = z'|x, z, do(x')) = \frac{2}{3}$. *We can compute that the resulting causal influence is* $\frac{3}{4} \log \frac{3}{2} \approx 0.43$.

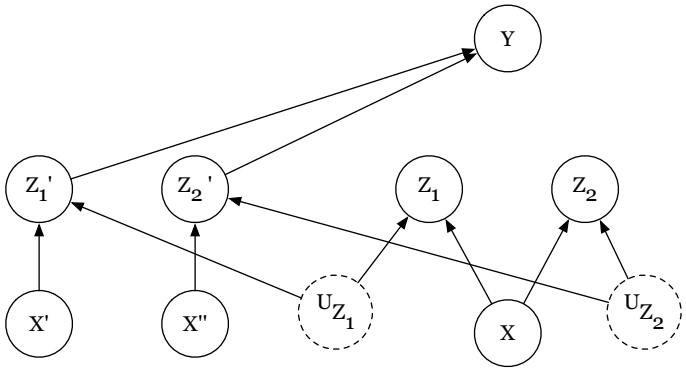
2.6 Sensitivity to Hidden vs. Observed Variables

The above example is also a good example of how changing a variable from hidden to observed or vice versa can change the indirect causal influence. In model 2 *above, if we change* U_Z *to an observed variable, the causal influence changes from* $\frac{3}{4} \log \frac{3}{2}$ *to* $\frac{1}{2}$. *This has the intuitive explanation that* X *completely controls* Y *with probability* $\frac{1}{2}$, *giving us a causal influence equal to* $\frac{1}{2}H(Y)$.

When U_Z *is unobserved and* $Z = X$, *the best we can do is infer that there is a* $\frac{2}{3}$ *chance that* $U_1 = 1$, *meaning that changing* X *only has a* $\frac{2}{3}$ *probability of affecting* Z. *As discussed in Section 1, when* U_Z *is hidden, we can think of ourselves as choosing* X *before* U_Z *is chosen; when* U_Z *is observed, we pick* X *after.*



(a)



(b)

Figure 4

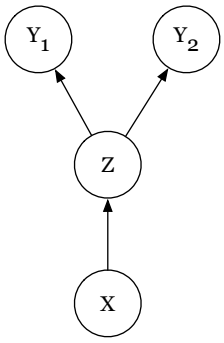
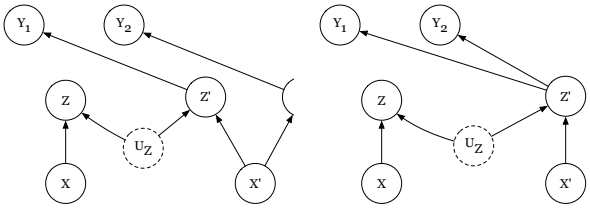


Figure 5



(a) (b)

Figure 6

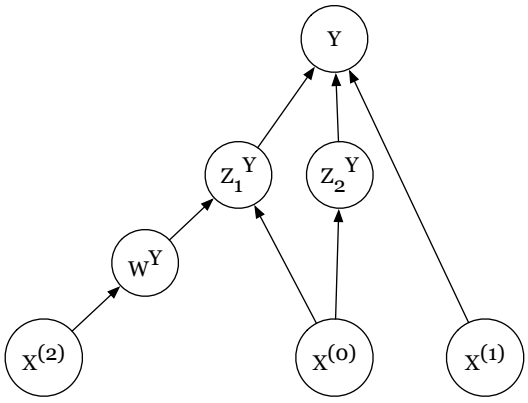


Figure 7

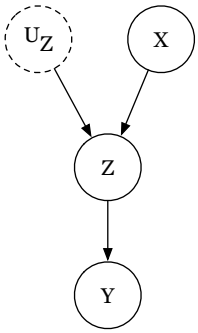


Figure 8

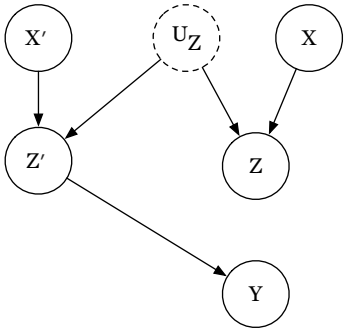


Figure 9