V13: Causality

Aims: (1) understand the causal relationships between the variables of a network
(2) interpret a Bayesian network as a causal model whose edges have causal significance.

For standard probabilistic queries it does not matter whether a Bayesian model is causal or not.

It matters only that it encode the „right“ distribution.

A correlation between 2 variables X and Y can arise in multiple settings:

- when X causes Y
- when Y causes X
- or when X and Y are both effects of a single cause.

Intervention queries

When some variable W causally affects both X and Y, we generally observe a correlation between them.

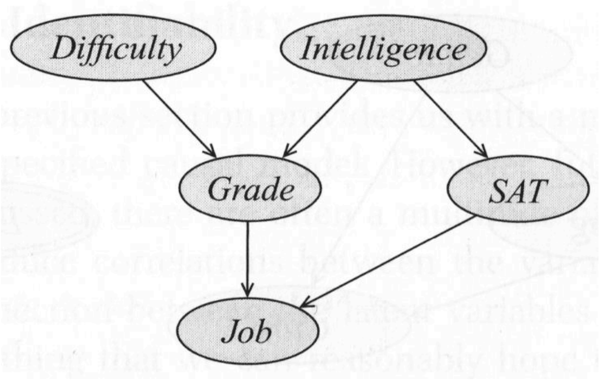
If we know about the existence of W and can observe it, we can disentangle the correlation between X and Y that is induced by W.

One approach to modeling causal relationships is using the notion of ideal **intervention** – interventions of the form do(Z := z), which force the variable Z to take the value z, and have no other immediate effect.

In **intervention queries**, we are interested in answering queries of the form P(Y | do(z)) or P(**Y** | do(z), **X** = **x**)

Case study: student example

Let us revisit our simple student example from lecture V8 and consider a particular student Gump.



Conditioning an observation that Gump receives an A in the class increases the probability that he has high intelligence, his probability of getting a high SAT score, and his probability of getting a job.

Consider a situation where Gump is lazy and rather than working hard to get an A in the class, he pays someone to hack into the university computer system and change his grade in the course to an A.

What is his probability of getting a good job in this case?

Intuitively, the company where Gump is applying only has access to his transcript.

Thus, we expect $P(J \mid do(g^1) = P(J \mid g^1)$.

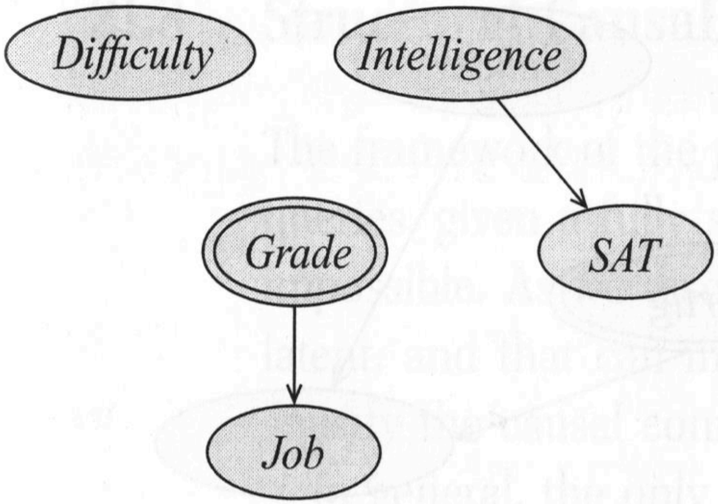
Case study: student example

What about the other 2 probabilities?

Intuitively, we feel that the manipulation to Gump‘s `grade` should not affect our beliefs about his `intelligence` nor about his `SAT score`.

Thus, we expect $P(i^1 \mid do(g^1)) = P(i^1)$ and $P(s^1 \mid do(g^1)) = P(s^1)$

An appropriate graphical model for the postintervention situation is shown in the right figure.



Here, the `grade` (and thus his chances to get the `job`) do no longer depend on `intelligence` or `difficulty` of the class.

This model is an instance of a **mutilated network**.

Latent variables

In practice, however, there is a huge set of possible **latent variables**, representing factors in the world that we cannot observe and often are not even aware of.

A latent variable may induce correlations between the observed variables that do not correspond to causal relations between them, and hence forms a **confounding factor** in our goal of determining causal interactions.

For the purpose of causal inference, it is critical to disentangle the component in the correlation between X and Y that is due to causal relationships and the component due to these confounding factors.

Unfortunately, it is virtually impossible in complex real-world settings, to identify all relevant latent variables and quantify their effects.

Causal models

A **causal model** has the same form as a probabilistic Bayesian network.

It consists of a directed acyclic graph over the random variables in the domain.

The model asserts that each variable X is governed by a causal mechanism that (stochastically) determines its value based on the values of its parents.

A causal mechanism takes the same form as a standard CPD.
For a node X and its parents **U**, the causal model specifies
for each value **u** of **U** a distribution over the values of X.

The difference to a probabilistic Bayesian network is in the interpretation of edges.

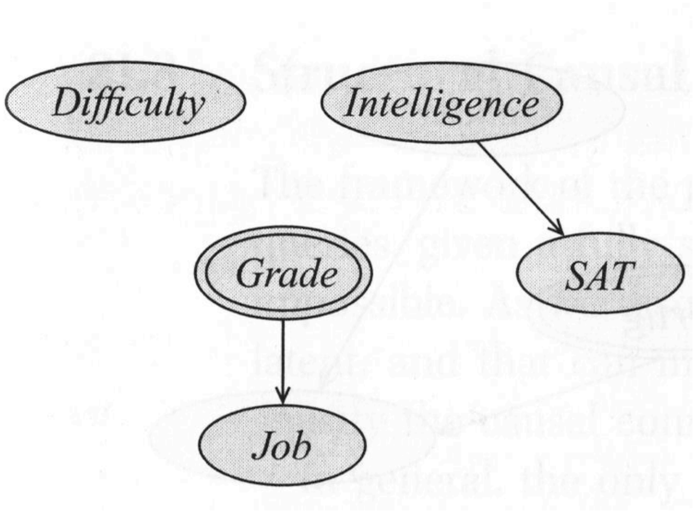
In a causal model, we assume that X‘s parents are its direct causes.

Causal models

The assumption that CPDs correspond to causal mechanisms forms the basis for the treatment of intervention queries.

Wen we intervene at a variable X, setting its value to x, we replace its original causal mechanism with one that dictates that it should take the value x.

The model presented before is an instance of the **mutilated network**.



In a mutilated network $B_{\mathbf{Z=z}}$, we eliminate all incoming edges into each variable $Z_i \in \mathbf{Z}$, (here: `grade`) and set its value to be z_i with probability 1.

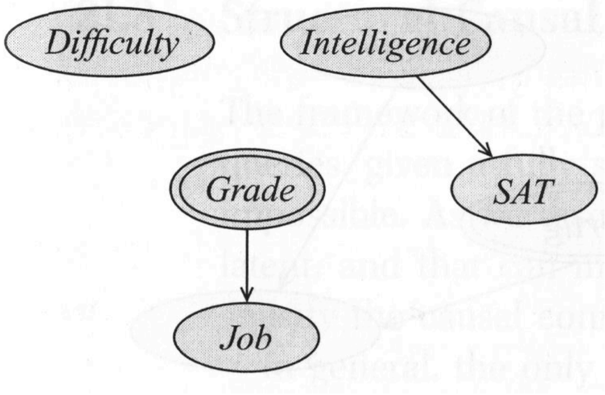
Causal models

Definition: A **causal model** C over X is a Bayesian network over X, which in addition to answering probabilistic queries, can also answer intervention queries P(**Y** | do(**z**), **x**) as follows:

$$P_C(\mathbf{Y} \mid do(\mathbf{z}), \mathbf{x}) = P_{C_{\mathbf{Z}=\mathbf{z}}}(\mathbf{Y} \mid \mathbf{x})$$

This approach deals appropriately with the student example.

Causal models



Let C^{student} be the appropriate causal model.

When we intervene in this model by setting Gump‘s grade to an A, we obtain the mutilated network shown before.

The distribution induced by this network over Gump‘s `SAT score` is the same as the prior distribution over his `SAT score` in the original network.

Thus as expected

$$P(S|do(G := g^1)) = P_{{C^{student}}_{G=g^1}}(S) = P_{C^{student}}(S)$$

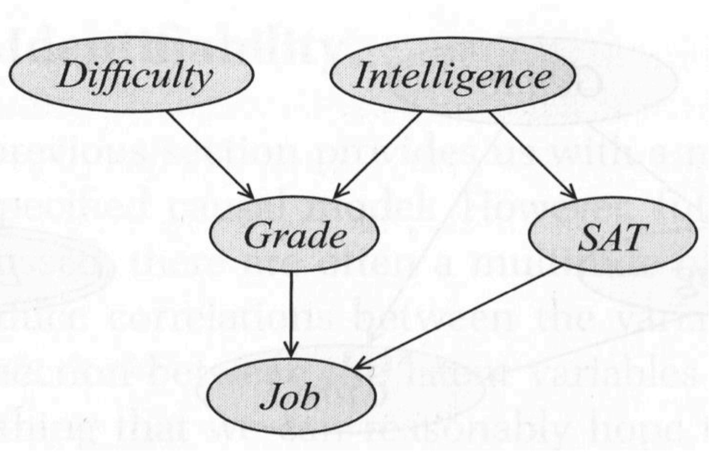
Conversely, the distribution induced on Gump‘s job prospects is

$$P_{C^{student}}(J \mid G = g^1)$$

Causal models

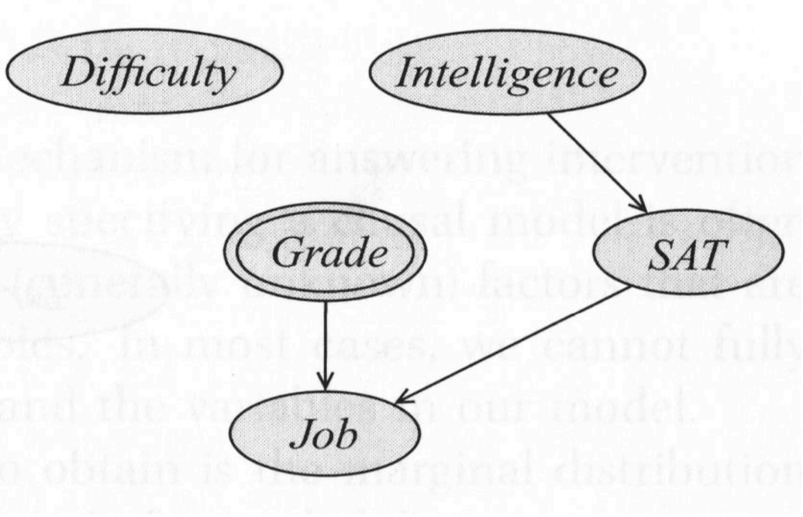
Assume that we start out with a somewhat different student model.

In this case, the recruiter can also base her hiring decision on the student‘s `SAT score`.



Now, the query $P_{C^{student}}(J \mid do(g^1))$ is answered by the mutilated network shown below.

The answer is clearly not $P_{C^{student}}(J)$ due to the direct causal influence of his grade on his job prospects.



It is also not equal to $P_{C^{student}}(J \mid G = g^1)$. because the new network also includes an influence via G ← I → S → J which is not present in the mutilated model.

Simpson‘s paradox

Consider the problem of trying to determine whether a **drug** is **beneficial** in curing a particular disease within some population of patients.

Statistics show that, within the population,

57.5% of patients who took the drug (D) are cured (C),

whereas only 50% of the patients who did not take the drug are cured.

One may belief given these statistics that the drug is beneficial.

However, within the subpopulation of **male patients**

70% who took the drug are cured, whereas

80% who did not take the drug are cured.

Within the subpopulation of **female patients**

20% of who took the drug are cured, whereas

40% of those who did not take the drug are cured.

Simpson‘s paradox

Thus, despite the apparently beneficial effect of the drug on the overall population, the drug appears to be detrimental to both men and women. We have

$$P(c^1|d^1) > P(c^1|d^0)$$
$$P(c^1|d^1, G = male) < P(c^1|d^0, G = male)$$
$$P(c^1|d^1, G = female) < P(c^1|d^0, G = female)$$

This surprising case can occur because taking this drug is correlated with gender.
In this particular example, 75% of men take the drug, but only 25% of women.

If the population contains 200 people, equally distributed into 100 men and 100 women,

75 men take the drug, of these 52.5 are cured (70%)
25 men did not take the drug, of these 20 are cured (80%).

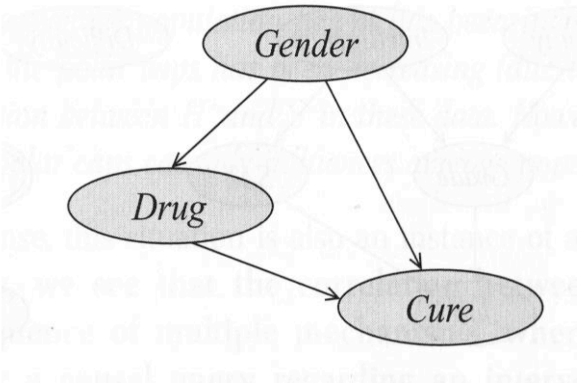
25 women take the drug, of these 5 are cured (20%)
75 women did not take the drug. Of these 30 are cured (40%).

→ of 100 people (men + women) taking the drug, 57.5 are cured (57.5%).
→ of 100 people not taking the drug, 50 are cured (50%).

Simpson‘s paradox

The conceptual difficulty behind this paradox is that it is not clear which statistics one should use when deciding whether to describe the drug to a patient.

The causal framework provides an answer to the problem on which variables we should condition on.



The appropriate query we need to answer in determining whether to prescribe the drug is not $P(c^1, d^1)$ but rather $P(c^1 \mid do(d^1))$.

We will show later that the correct answer is that the drug is not beneficial, as expected.

Structural Causal Identifiability

Fully specifying a causal model is often impossible.

At least, for intervention queries, we must disentangle the causal influence of X and Y from other factors leading to correlations between them.

Structural Causal Identifiability

Consider a pair of variables X, Y with an observed correlation between them, and imagine that our task is to determine P(Y | do(X)).

Let us even assume that X temporally precedes Y,
and therefore we know that Y cannot cause X.

However, if we consider the possibility that least some of the correlation between X and Y is due to a hidden common cause,
we have no way of determining how much effect perturbing X would have on Y.

If all of the correlation is due to a causal link, then P(Y | do(X)) = P(Y | X).

Conversely, if all of the common correlation is due to the hidden common cause, then P(Y | do(X)) = P(Y).

In general, any value between those 2 distributions is possible.

Query Simplification Rules

Which intervention queries are identifiable?

We will now see that the structure of a causal model
gives rise to certain **equivalence rules** over interventional queries.

This allows one query to be replaced by an equivalent one
that may have a simpler form.

These rules can be defined in terms of an **augmented causal model**
that encodes the possible effect of interventions explicitly
within the graph structure.

We will view the process of an intervention in terms
of a new **decision variable** $\hat{Z}$ that determines
whether we intervene at Z, and if so, what its value is.

Query Simplification Rules

The variable $\hat{Z}$ takes on values in $\{\epsilon\} \cup Val(Z)$.

If $\hat{Z} = \epsilon$, then Z behaves as a random variable whose distribution is determined by its usual CPD $P(Z|Pa_Z)$.

If $\hat{Z} = z$, then it deterministically sets the value of Z to be z with probability 1.

Let $\hat{\mathbf{Z}}$ denote the set $\{\hat{Z}: Z \in \mathbf{Z}\}$

In those cases, where Z‘s value is deterministically set by one parent, all of Z‘s other parents **U** become irrelevant so that their edges to Z can be removed.

Let $\boldsymbol{G}^{\dagger}$ be the augmented model for G. Let $\boldsymbol{G}^{\dagger}{}_{\overline{\mathbf{Z}}}$ be the graph obtained from $G^{\dagger}$ except that every $Z \in \mathbf{Z}$ has only the single parent $\hat{Z}$.

Query Simplification Rules

The first query simplification rule allows us
to insert or delete observations into a query.

Proposition. Let C be a causal model over the graph structure G. Then:

$$P(\mathbf{Y} \mid do(\mathbf{Z} \coloneqq \mathbf{z}), \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}) = P(\mathbf{Y} \mid do(\mathbf{Z} \coloneqq \mathbf{z}), \mathbf{X} = \mathbf{x})$$

if **W** is d-separated from **Y** given **Z**, **X** in the graph $G^{\dagger}{}_{\bar{Z}}$.

Nomenclature:
We say that **X** and **Y** are **d-separated** given **Z**,
if there is no active trail between any node
X ∈ **X** and Y ∈ **Y** given **Z**.

Query Simplification Rules

The second rule is subtler and allows us
to replace an intervention with the corresponding observation.

Proposition. Let C be a causal model over the graph structure G.Then:

$$P(\mathbf{Y} \mid do(\mathbf{Z} \coloneqq \mathbf{z}), do(\mathbf{X} = \mathbf{x}), \mathbf{W} = \mathbf{w}) = P(\mathbf{Y} \mid do(\mathbf{Z} \coloneqq \mathbf{z}), \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$$

if **Y** is d-separated from $\widehat{\mathbf{X}}$ given **X**, **Z, W** in the graph $G^{\dagger}{}_{\bar{Z}}$.

This rule holds because it tells us that we do not
get more information regarding **Y** from the fact
that an intervention took place at **X** than the values **x** themselves.

Query Simplification Rules

The third rule allows us to introduce or delete interventions.

Proposition. Let C be a causal model over the graph structure G.Then:

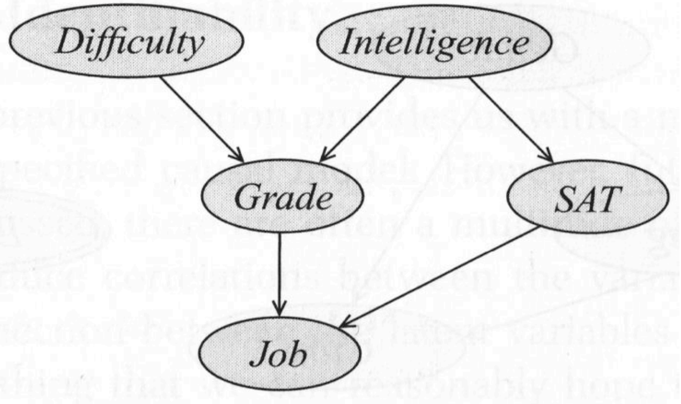
$$P(\mathbf{Y} \,|\, do(\mathbf{Z} \coloneqq \mathbf{z}), do(\mathbf{X} = \mathbf{x}), \mathbf{W} = \mathbf{w}) = P(\mathbf{Y} \,|\, do(\mathbf{Z} \coloneqq \mathbf{z}), \mathbf{W} = \mathbf{w})$$

if **Y** is d-separated from $\widehat{\mathbf{X}}$ given **Z, W** in the graph $G^{\dagger}{}_{\bar{Z}}$.

Iterated Query Simplification

There are many queries where none of the 3 rules apply directly.

But we will see that we can also perform other transformations on the query allowing the rules to be applied.



Let us revisit the right figure
which involves the query P(J | do(G)).
None of our rules apply directly to this query.

- we cannot eliminate the intervention as P(J | do(G)) ≠ P(J)

- we also cannot turn the intervention into an observation, P(J | do(G)) ≠ P(J | G)

because intervening at G only affects J via the single edge G → J, whereas conditioning G also influences J by the indirect trail G ← I → S → J.
This trail is called **a back-door trail**, since it leaves G via the „back door".

Iterated Query Simplification

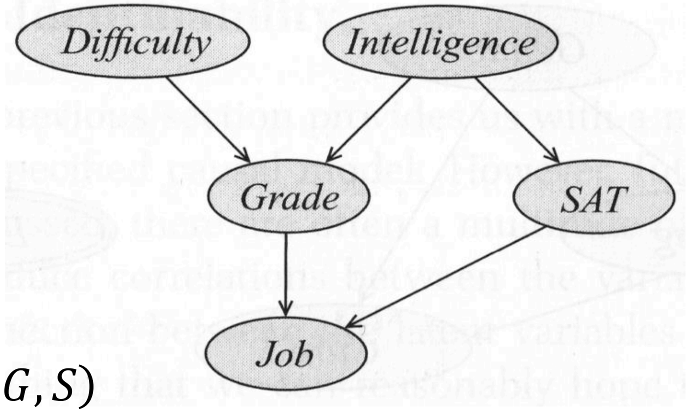
However we can use standard probabilistic reasoning and obtain:

$$P(J \mid do(G)) = \sum_S P(J \mid do(G), S) P(S \mid do(G))$$

Both of the terms in the summation can be further simplified.

For the first term, we have that the only active trail from G to J is the direct edge G → J.

I.e. J is d-separated from G given S in the graph where outgoing arcs from G have been deleted.



Thus we can apply the second rule and conclude

$$P(J \mid do(G), S) = P(J \mid G, S)$$

Iterated Query Simplification

For the second term $P(S \mid do(G))$, we already argued $P(S \mid do(G)) = P(S)$
(hacking the computer system does not change our belief on his intelligence/SAT score).

Putting the two together yields $P(J \mid do(G)) = \sum_S P(J \mid G, S)P(S)$
thus hacking the computer system does not affect the job chances anymore.

A **back-door trail** from X to Y is an active trail that leaves X via a parent of X.

For a query $P(\mathbf{Y} \mid do(\mathbf{X}))$, a set **W** satisfies the back-door criterion if no node in **W** is a descendant of **X** , and **W** blocks all back-door paths from **X** to **Y**.

One can show that if a set **W** satisfies the back-door criterion
for a query $P(\mathbf{Y} \mid do(\mathbf{X}))$, then

$$P(\mathbf{Y} \mid do(\mathbf{X})) = \sum_{\mathbf{W}} P(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) P(\mathbf{W})$$

Revisit Simpson‘s paradox

We will reconsider Simpson‘s paradox using the back-door criterion.

Consider again the query $P(c^1 \mid do(d^1))$.

The variable G (`gender`) introduces a back-door trail between C and D.

We can account for its influence using the eq. just derived:

$$P(c^1 \mid do(d^1)) = \sum_g P\,(c^1 \mid d^1, g)P(g).$$

we obtain:

$$P(c^1 \mid do(d^1)) = 0.7 \cdot 0.5 + 0.2 \cdot 0.5 = 0.45$$
$$P(c^1 \mid do(d^0)) = 0.8 \cdot 0.5 + 0.4 \cdot 0.5 = 0.6$$

Therefore, we should not prescribe the drug.

Case study: lung cancer

In the early 1960s, following a significant increase in the number of smokers that occurred around World War II, people began to notice a substantial increase in the number of cases of lung cancer.

After many studies, a **correlation** was noticed between **smoking** and **lung cancer.**

This correlation was noticed in both directions:
The frequency of smokers among lung cancer patients was substantially higher than in the general population,

Also, the frequency of lung cancer patients within the population of smokers was substantially higher than within the population of nonsmokers.

This led the Surgeon General, in 1964, to issue a report linking cigarette smoking to lung cancer.

This report came under severe attack by the tobacco industry.

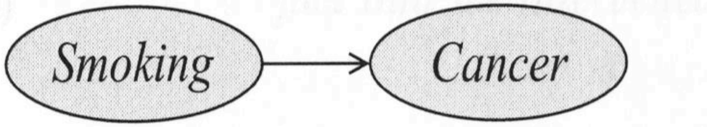
Case study: lung cancer

The industry claimed that the observed correlation can also be explained by a model in which there is no causal relationship between `smoking` and `lung cancer`.

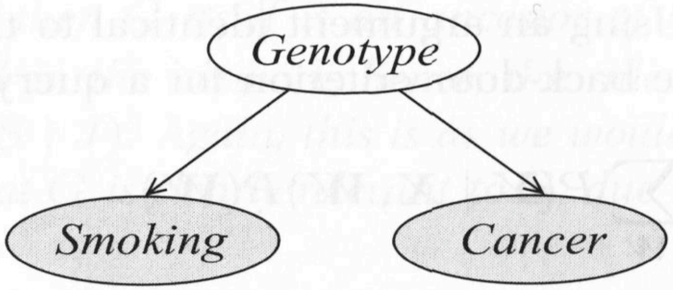
Instead, an observed `genotype` might exist that simultaneously causes cancer and a desire for nicotine.

There exist several possible models

Direct causal effect



Indirect influence via a latent common parent genotype



Case study: lung cancer

The 2 models can express precisely the same set of distributions over the observable variables S and C.

Thus, they can do an equally good job of representing the empirical distribution over these variables, and there is no way to distinguish between them based on observational data alone.

Both models will provide the same answer to standard probabilistic queries such as $P(c^1 \mid s^1)$.

However, relative to interventional queries, these models have very different consequences.

Case study: lung cancer

According to the Surgeon General‘s model, we would have

$$P(c^1 \mid do(S \coloneqq s^1)) = P(c^1 \mid s)$$

In other words, if we force people to smoke, their probability of getting cancer is the same as the probability conditioned on smoking, which is much higher than the prior probability.

According to the tobacco industry model, we have

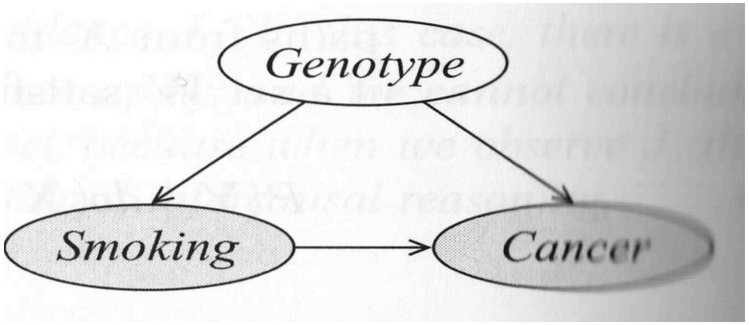
$$P(c^1 \mid do(S \coloneqq s^1)) = P(c^1)$$

In other words, making the population smoke or stop smoking would have no effect on the rate of cancer cases.

Pearl (1995) proposed a formal analysis of this dilemma.
He proposed that we should combine these 2 models into a single joint model.

Case study: lung cancer

Pearl model



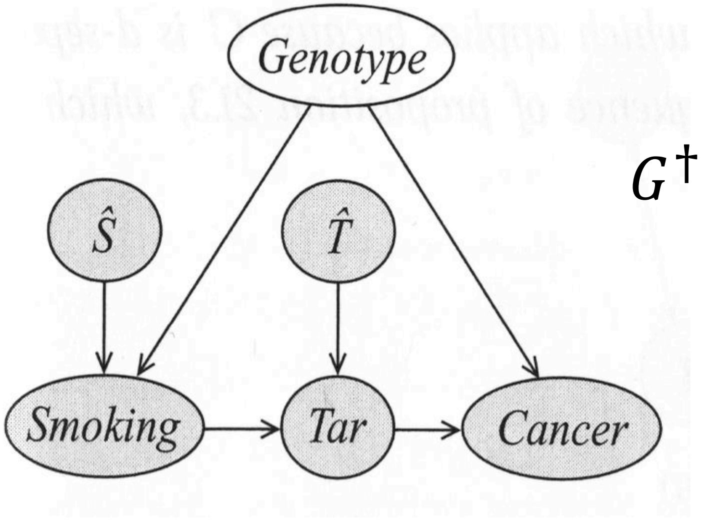
We now need to assess from the marginal distribution over the observed variables alone the parametrization of the 3 links.

Unfortunately, it is impossible to determine the parameters of these links from the observational data alone, since both original models can explain the data perfectly.

Pearl thus refined the model somewhat by introducing an additional assumption, and could then determine estimates for the links.

Case study: lung cancer

Assume that we determine that the effect of smoking on cancer is not a direct effect, but occurs through the accumulation of `tar` deposits in the lungs, see figure.



Here, we assume that the deposition of tar in the lungs is not directly affected by the latent Genotype variable.

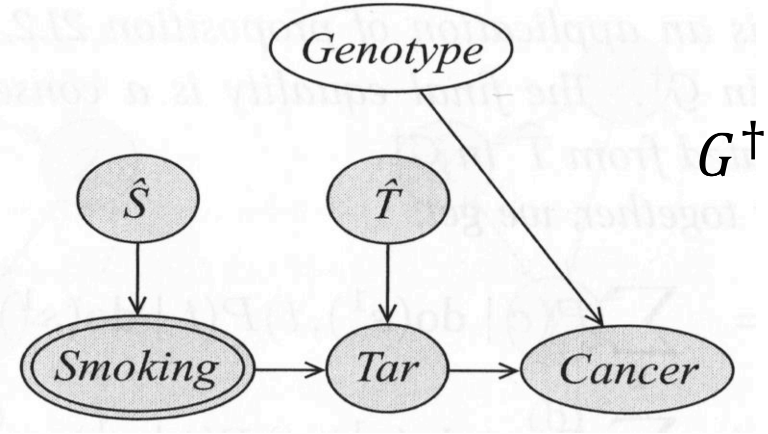
We will now show that, if we can measure the amount of tar deposits in the lungs of various individuals (e.g. by X-ray or in autopsies), we can determine the probability of the intervention query

$$P(c^1|do(s^1))$$

using observed correlations alone.

Case study: lung cancer

We are interested in $P(c^1 \mid do(s^1))$, which is an intervention query whose mutilated network is



$G^{\dagger}{}_{\bar{S}}$

Standard probabilistic reasoning shows that

$$P(C \mid do(s^1)) = \sum_t P(C \mid do(s^1), t) P(t \mid do(s^1))$$

We now consider and simplify each term separately.

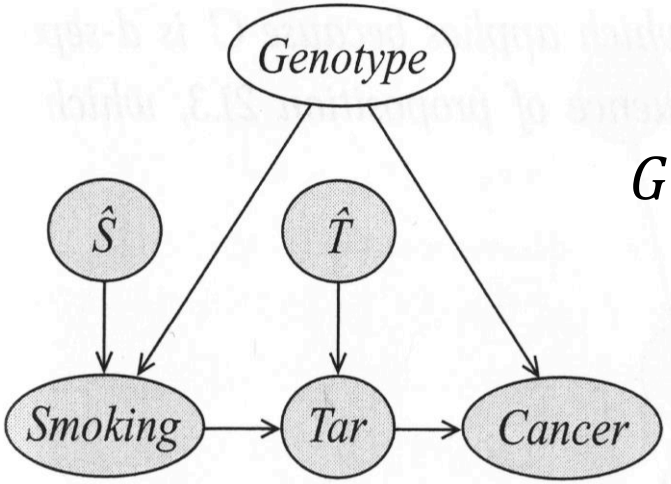
Case study: lung cancer

The second term, which measures the effect of `smoking` on `tar`, can be simplified directly using our rule for converting interventions to observations (second rule).

Here, $\hat{S}$ is d-separated from T given S in graph $G^{\dagger}$.

It follows that

$$P(t \mid do(s)) = P(t \mid s)$$



$G^{\dagger}$

Intuitively, the only active trail from $\hat{S}$ to T goes via S, and the effect of that trail is identical regardless of whether we condition on S or intervene at S.

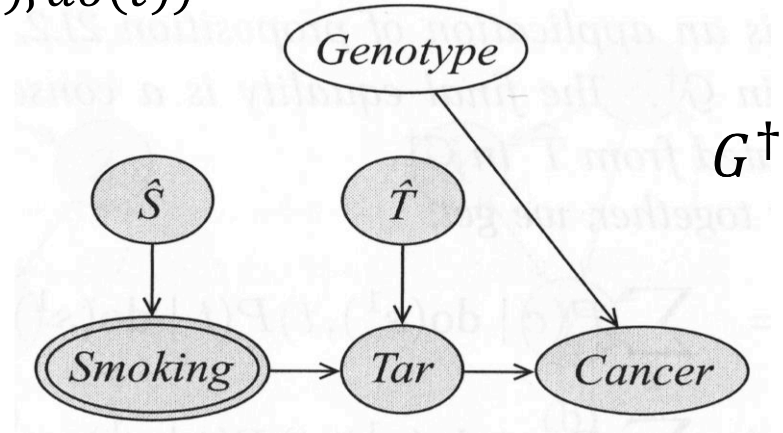
Case study: lung cancer

The first term $P(C \mid do(s^1), t)$ measures the effect of `tar` on `cancer` in the presence of our intervention on S.

Unfortunately, we cannot directly convert the intervention at S to an observation, since C is not d-separated from $\hat{S}$ given S, T in $G^{\dagger}$.

However, we can convert the observation at T to an intervention, because C is d-separated from $\hat{T}$ given S,T in the graph $G^{\dagger}{}_{\bar{S}}$.

$$P(C \mid do(s^1), t) = P(C \mid do(s^1), do(t))$$



$G^{\dagger}{}_{\bar{S}}$

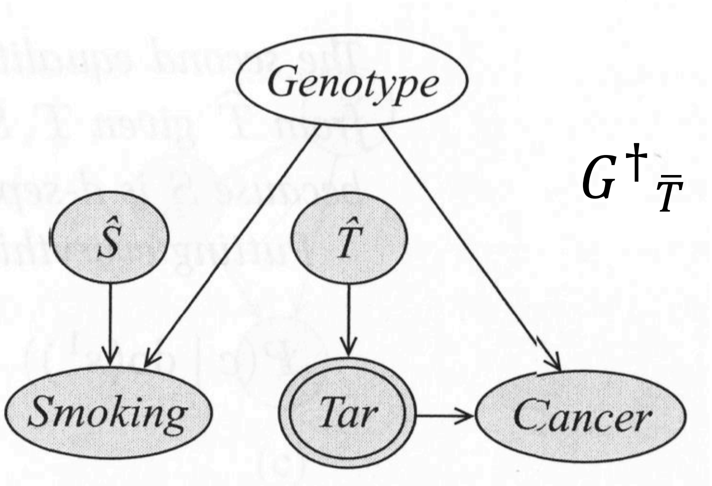
Case study: lung cancer

We can now eliminate the intervention at S from this expression using the third rule, which applies because C is d-separated from $\hat{S}$ given T in the graph $G^{\dagger}{}_{\bar{T}}$.

We obtain

$$P(C \mid do(s^1), do(t)) = P(C \mid do(t))$$

By standard probabilistic reasoning and conditioning on S we get



$$P(C \mid do(t)) = \sum_{s'} P(C \mid do(t), s')P(s' \mid do(t)) =$$

By rule 2 because C is d-separated from $\hat{T}$ given T, S in $G^{\dagger}$

$$\sum_{s'} P(C \mid t, s')P(s' \mid do(t)) =$$

By rule 3 because S is d-separated from $\hat{T}$ in $G^{\dagger}$

$$\sum_{s'} P(C \mid t, s')P(s')$$

Case study: lung cancer

Putting everything together, we get

$$P(c \mid do(s^1)) = \sum_t P(c \mid do(s^1), t) P(t \mid do(s^1))$$

$$= \sum_t P(c \mid do(s^1), t) P(t \mid s^1)$$

$$= \sum_t P(t \mid s^1) \sum_{s'} P(c \mid t, s') P(s')$$

Thus, if we agree that tar in the lungs is the intermediary between smoking and lung cancer,
we can uniquely determine the extent to which smoking causes lung cancer
even in the presence of a confounding latent variable (here: `genotype`).