# GT Causalité

**Séance 2**

This session is based on Brady Neal's lecture notes (Chapter 3, and part of Chapter 4). It also uses [KF09, Chapter 3], [PJS17, Chapters 2 and 6], and [Pea22, Chapters 1 and 2]. Sometimes I refer to [Lau96, Chapter 3] for some proofs.

## 1 Bayesian networks

Here I note that the explanations in Brady Neal's lecture notes are quite cryptic, so I am mostly following the story in [KF09]. Something that does not help in Neal's notes is that he amalgamates causality concepts with bayesian networks terminology which are unrelated with causality. Indeed, in this section we shall not think at all about causality and only think about independence relations.

To make things a bit more simple, here we consider random variables $(X_1, \dots, X_d) \in \mathcal{X}_1 \times \dots \mathcal{X}_d$ with distribution $P$ and we do assume that $P$ admits a density $p$ with respect to the product measure $\mu_1 \times \cdots \times \mu_d$.

We shall use the following vocabulary about DAGs: $\mathrm{Pa}(X)$ denote the set of *parents* of $X$, *ie.* nodes $Y$ such that there is a directed edges $Y \to X$; $\mathrm{NonDesc}(X)$ the set of non descendants of $X$, *ie.* nodes in $V \backslash X$ that cannot be reached from $X$ through a directed path.

**Definition 1.** *Let $G$ be a DAG on vertices $V = \{X_1, \dots, X_d\}$. Then $P$ factorizes according to $G$ is there are conditional densities $p_i : \mathcal{X}_i \times \prod_{j \in \mathrm{Pa}(X_i)} \mathcal{X}_j \to \mathbb{R}_+$ such that*

$$p(x_1, \dots, x_n) = \prod_{i=1}^{d} p_i(x_i \mid (x_j)_{j \in \mathrm{Pa}(X_i)}) \qquad P - \text{as.} \tag{1}$$

**Definition 2.** *A bayesian network is a pair $(G, P)$ where $G$ is a DAG and $P$ factorizes according to $G$.*

**Example 1.** *Consider the DAG $X_1 \to X_2 \to X_3$. The distribution $P$ with density $p(x_1, x_2, x_3) = p_1(x_1) p_{2|1}(x_2 \mid x_1) p_{3|2}(x_3 \mid x_2)$ factorizes according to the DAG.*

The graph $G$ can be viewed in two very different ways:

1. as a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way;

2. as a compact representation for a set of conditional independence assumptions about a distribution.

As we will see, these two views are, in a strong sense, equivalent.

Given three subsets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subset V$ we say that $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ holds in $P$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given $\boldsymbol{Z}$; if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent unconditionally, we say $(\boldsymbol{Y} \perp \boldsymbol{Z})$ or $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \varnothing)$.

**Definition 3.** *The set of all independence relations of $P$, written $\mathcal{I}(P)$ is defined as the set of relations $(A \perp B \mid C)$ that hold in $P$.*

As said in Item 2, we will see that the statement "$P$ factorizes according to $G$" is indeed equivalent to some statements about $\mathcal{I}(P)$. More precisely, we shall see (eventually under mild positivity conditions on $p$) that Definition 1 is equivalent to the following two definitions:

**Definition 4** (Local Markov property)**.** *$P$ satisfies the local Markov property with respect to $G$ iff*

$$\mathcal{I}_{\text{loc}}(G) \coloneqq \{(X_i \perp \text{NonDesc}(X_i) \mid \text{Pa}(X_i)) \,:\, i = 1, \dots, d\} \subset \mathcal{I}(P).$$

**Definition 5** (Global Markov property)**.** *$P$ satisfies the global Markov property with respect to $G$ iff*

$$\mathcal{I}(G) \coloneqq \{(\boldsymbol{X}, \boldsymbol{Y} \mid Z) \,:\, \text{d-sep}_G(\boldsymbol{X}, Y \mid Z)\} \subset \mathcal{I}(P).$$

To fully understand the global Markov property, we must first define d-separation, which will be addressed later; for now, we just point that the set $\mathcal{I}(G)$ is richer than the set $\mathcal{I}_{loc}(G)$ and typically contains more independence relations.

## 1.1 Local Markov property and factorization

We wish to show that $\mathcal{I}_{\text{loc}}(G) \subset \mathcal{I}(P)$ (*aka. local Markov property*) iff $P$ factorizes according to $G$.

**Theorem 1.** *The following statements are equivalent:*

*1. $P$ factorizes according to $G$;*

*2. $P$ satisfies the local Markov property relative to $G$ (ie. $\mathcal{I}_{\text{loc}}(G) \subset \mathcal{I}(P)$).*

*Proof.* (1) $\implies$ (2) See [KF09, Theorem 3.2] or [Lau96, Section 3.2].

(2) $\implies$ (1) [KF09, Theorem 3.1] or [Lau96, Section 3.2]. □

**Example 2.** *Consider $A \to B \to C$, $A \to D$ $E$. Then $(C \perp \{A, D, E\} \mid B)$ is a local independence relation. On the other hand, if $P$ factorizes according to this graph, then $(A \perp \{C, E\} \mid \{B, D\})$ which is non-local (but we will is captured by the global Markov property and d-separation).*

## 1.2 Global Markov property, $d$-separation

As we discussed, a graph structure $G$ encodes a certain set of conditional independence assumptions $\mathcal{I}_{\mathrm{loc}}(G)$. Knowing only that a distribution $P$ factorizes over $G$, we can conclude that it satisfies the local Markov property. An immediate question is whether there are other independencies that we can "read-off" directly from $G$. That is, are there other independencies that hold for every distribution $P$ that factorizes over $G$?

### 1.2.1 Basic building blocks of $G$ and intuitions

Here we take inspiration from Brady Neal's sections 3.5 and 3.6; our goal is to understand the ideas motivating the definition of $d$-separation.

We use the example to define various motifs of interest, and give a hint on why they are interesting.

**Example 3** (Chains). *A chain is a motif of the form $X_1 \to X_2 \to X_3$. Consider the graph on 3 vertices which is a chain and $P$ factorizing according to it (equivalently $P$ satisfies the local Markov property relative to $G$). Usually $X_1$ and $X_3$ are dependent. But if we look at $X_1, X_3 \mid X_2$, we can see that*

$$p_{123}(x_1, x_2, x_3) = p_1(x_1) p_{2|1}(x_2 \mid x_1) p_{3|2}(x_3 \mid x_2) \tag{2}$$

*so that $X_1, X_3 \mid X_2 = x_2$ admits the density*

$$\begin{aligned} p_{13|2}(x_1, x_2 \mid x_2) &= \frac{p_{123}(x_1, x_2, x_3)}{p_2(x_2)} && (3) \\ &= \frac{p_1(x_1) p_{2|1}(x_2 \mid x_1)}{p_2(x_2)} p_{3|2}(x_3 \mid x_2) && (4) \\ &= p_{13|2}(x_1, x_3 \mid x_2) p_{3|2}(x_3 \mid x_2). && (5) \end{aligned}$$

*In other words, $(X_1 \perp X_3 \mid X_2)$ holds in $P$. We shall say that $X_2$ blocks the path between $X_1 \to X_2 \to X_3$.*

**Example 4** (Forks). *A fork is a motif of the form $X_1 \leftarrow X_2 \to X_3$. It is easily seen that forks and chains encodes the same type of conditional independencies.*

**Example 5** (Colliders and immoralities). *A collider is a motif of the form $X_1 \to X_2 \leftarrow X_3$, if there is no edge between $X_1$ and $X_3$, the motif is called an immorality[1]. Consider the graph on 3 vertices which is an immorality and $P$ factorizing according to it. Then,*

[1] The term "immoral" is used humorously to indicate that these parents have not formed a relationship, even though they share a child.

*we can see that $X_1 \perp X_3$ holds in $P$ because*

$$p_{13}(x_1, x_3) = \int p_{123}(x_1, x_2, x_3) \mathrm{d}\mu_2(x_2) \tag{6}$$
$$= \int p_{123}(x_1, x_2, x_3) \mathrm{d}\mu_2(x_2) \tag{7}$$
$$= \int p_1(x_1) p_3(x_3) p_{2|13}(x_2 \mid x_1, x_3) \mathrm{d}\mu_2(x_2) \tag{8}$$
$$= p_1(x_1) p_3(x_3) \int p_{2|13}(x_2 \mid x_1, x_3) \mathrm{d}\mu_2(x_2) \tag{9}$$
$$= p_1(x_1) p_3(x_3). \tag{10}$$

*But, if we look at $X_1, X_3 \mid X_2$, then*

$$p_{13|2}(x_1, x_3 \mid x_2) = \frac{p_{123}(x_1, x_2, x_3)}{p_2(x_2)} = p_1(x_1) p_2(x_2) \frac{p_{2|13}(x_2 \mid x_1, x_3)}{p_2(x_2)}. \tag{11}$$

*So oddly-enough, conditional on $X_2$ the variables $X_1$ and $X_3$ may becomes dependent if $X_2$ is really depending on $(X_1, X_3)$. Brady Neal has the following "concrete" example: An example is the easiest way to see why this is the case. Imagine that you're out dating men, and you notice that most of the nice men you meet are not very good-looking, and most of the good-looking men you meet are jerks. It seems that you have to choose between looks and kindness. In other words, it seems like kindness and looks are negatively associated. However, what if I also told you that there is an important third variable here: availability (whether men are already in a relationship or not)? And what if I told you that a man's availability is largely determined by their looks and kindness; if they are both good-looking and kind, then they are in a relationship. The available men are the remaining ones, the ones who are either not good-looking or not kind. You see an association between looks and kindness because you've conditioned on a collider (availability). You're only looking at men who are not in a relationship. The structure of this example is* looks $\rightarrow$ availability $\leftarrow$ kindness.

**Example 6** (Descendant of immoralities)**.** *Similarly, conditionning on a descendants of an immorality can induce association in between the parents of the collider. The intuition is that if we learn something about a collider's descendant, we usually also learn something about the collider itself because there is a direct causal path from the collider to its descendants, and we know that nodes in a chain are eventually dependent.*

What is important to retain here is that conditionning on the middle vertex of a chain or a fork can "block" the flow of dependencies along it. In contrast, conditionning on the middle vertex of an immorality can "unblock" the flow.

### 1.2.2 Paths, blocked paths, and $d$-separation

A (undirected) path is a sequence of distinct vertices $(X_{i_1}, \ldots, X_{i_m})$, $m \geq 2$, such that for each $k = 1, \ldots, m-1$ there is an edge $X_{i_k} \rightarrow X_{i_{k+1}}$ or $X_{i_k} \leftarrow X_{i_{k+1}}$.

**Definition 6** (Blocked path)**.** *A (undirected) path $(X_{i_1}, \ldots, X_{i_m})$ is blocked by a set of vertices $\boldsymbol{Z}$ (not containing $X_{i_1}$ nor $X_{i_m}$) if and only if:*

- $(X_{i_1}, \ldots, X_{i_m})$ *contains a chain* $X_{i_{k-1}} \to X_{i_k} \to X_{i_{k+1}}$ *or a fork* $X_{i_{k-1}} \leftarrow X_{i_k} \to X_{i_k}$ *such that* $X_{i_k} \in \boldsymbol{Z}$*, or*

- $(X_{i_1}, \ldots, X_{i_m})$ *contains an immorality* $X_{i_{k-1}} \to X_{i_k} \leftarrow X_{i_{k+1}}$ *such that* $X_{i_k} \notin \boldsymbol{Z}$ *and no descendant of* $X_{i_k}$ *is in* $\boldsymbol{Z}$*.*

[draw picture! in particular compare the difference between immorality and a collider $X \to Y \leftarrow Z$ with an edge between $X$ and $Z$]

**Definition 7** (d-separation)**.** *Let* $\boldsymbol{X}$*,* $\boldsymbol{Y}$ *and* $\boldsymbol{Z}$ *be three disjoint subsets of vertices.* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are d-separated by* $\boldsymbol{Z}$ *if every path between vertices of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *is blocked by* $\boldsymbol{Z}$*. We then write*

$$\text{d-sep}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}).$$

### 1.2.3 The global Markov property

Recall $\mathcal{I}(G) \coloneqq \{(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) : \text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})\} \subset \mathcal{I}(P)$.

**Theorem 2.** *If* $p > 0$ *the following are equivalent:*

1. *$P$ factorizes according to $G$;*

2. *$P$ satifies the global Markov property relative to $G$ (ie. $\mathcal{I}(G) \subset \mathcal{I}(P)$).*

*Proof.* (1) $\implies$ (2) [Lau96, Corollary 3.23].

(1) $\implies$ (2) It is enough to observe that for every vertex $X_i$ the sets $\{X_i\}$ and $\mathrm{NonDesc}(X_i)$ are d-separated by $\mathrm{Pa}(X_i)$. In other words the global Markov property implies the local Markov property, which in turn implies (1) by Theorem 1. $\square$

So indeed, we have equivalence between factorization, local Markov, and global Markov properties. This tells us that in a Bayesian network, the DAG gives indication about the independence relations in $\mathcal{I}(G)$ (which might not be all of $\mathcal{I}(P)$, see below).

To make it interesting, however, we shall ensure that the global Markov condition is indeed stronger than the local Markov conditions. This can be seen because the parents of a node always d-separate the node from its non-descendants.

Interestingly, there are efficient algorithms to find out if two sets are $d$-separated given a third set [KF09, Algorithm 3.1]; so given the DAG we can immediately test if some independence relation is in $\mathcal{I}(G)$.

## 1.3 Faithfulness, aka converse global Markov property

**Definition 8** (Faithfulness)**.** *$P$ is faithful to $G$ if $\mathcal{I}(P) \subset \mathcal{I}(G)$.*

It is interesting to wonder if factorization implies faithfulness. This is because since factorization implies global Markov, then we would have that factorization implies $\mathcal{I}(G) = \mathcal{I}(P)$ and be happy to have encoded in the DAG all the independence relations of $P$. Unfortunately, factorization does not imply faithfulness, as seen in the examples below:

**Example 7.** *Suppose $X_1 \perp X_2$. Yet, the law of $(X_1, X_2)$ factorizes according to the graph $X_1 \to X_2$, but $\{X_1\}$ and $\{X_2\}$ are not d-separated, so we cannot read the independence relation from the DAG $X_1 \to X_2$.*

**Example 8.** *See also [PJS17, Example 6.34].*

This shows that in general we cannot read all the independence from the DAG in a Bayesian network. But, faithfulness is an important concept since, factorization + faithfulness implies that $\mathcal{I}(G) = \mathcal{I}(P)$; in other words that $d$-separation permits to characterize all independence relations.

## 1.4 Completeness of $d$-separation

The previous section shows that there can exist independence relations in $\mathcal{I}(P)$ that we cannot read from $\mathcal{I}(G)$; ie. some independence relations cannot be uncovered using $d$-separation. It is natural to ask whether or not $d$-separation is the best we can do. Indeed it is.

**Theorem 3** (Completeness)**.** *If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not $d$-separated given $\boldsymbol{Z}$ in $G$, then there exists a distribution $P$ that factorizes according to $G$ and in which $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent.*

*Proof.* [KF09, Theorem 3.4] □

We can view the completeness result as telling us that our definition of $\mathcal{I}(G)$ is the maximal one. For any independence relation n that is not a consequence of $d$-separation in $G$, we can always find a counterexample distribution $P$ that factorizes over $G$.

## 1.5 Markov equivalence

**Definition 9.** *Two graphs $G_1$ and $G_2$ are Markov equivalent if $\mathcal{I}(G_1) = \mathcal{I}(G_2)$.*

**Example 9** (Chains and forks)**.** *$X_1 \to X_2 \to X_3$, $X_1 \leftarrow X_2 \leftarrow X_3$, and $X_1 \leftarrow X_3 \to X_2$ are Markov equivalent (notice the importance for causality: in the causal interpretation, they are very different graphs since they don't suppose the same cause-effects relations, but probabilistically, they represent the same set of independence relations).*

*So the previous graphs are Markov equivalent, but they are not equivalent to the immorality $X_1 \to X_2 \leftarrow X_3$. This is because the immorality implies that $X_1 \perp X_3$ (see for instance Example 5), but there are distributions $P$ that factorizes according to chains or forks in which $X_1$ and $X_3$ are dependent.*

**Theorem 4.** *Two DAG $G_1$ and $G_2$ are Markov equivalent if and only if they have the same skeleton and the same set of immoralities.*

*Proof.* [KF09, Section 3.3.4]. □

## 1.6 Minimality

**Definition 10** (Minimality)**.** *If $P$ factorizes according to $G$, but doesn't factorize according to any proper sugraph of $G$, then $G$ is minimal.*

This also can be understood has: if we were to remove any edges from the DAG, $P$ would factorizes according to the graph with the removed edges.

**Theorem 5.** *For every $P$ there exists a minimal DAG.*

*Proof.* Let us exhibit such a DAG. Starting point is to notice that we can always write

$$p(x_1, \dots, x_d) = p_1(x_1) p_{2|1}(x_2 \mid x_1) \dots p_{d|1\dots(d-1)}(x_d \mid x_1, \dots, x_{d-1}).$$

Let $\mathrm{Pa}_1 = \varnothing$. For each $j = 2, \dots, d$ find the minimal subset $\mathrm{Pa}_j$ of $\{1, \dots, j-1\}$ for which

$$p_{j|1\dots(j-1)}(X_j \mid X_1, \dots X_{j-1}) = p_{j|\mathrm{Pa}_j}(X_j \mid (X_k)_{k \in \mathrm{Pa}_j}) \qquad P\text{-as.}$$

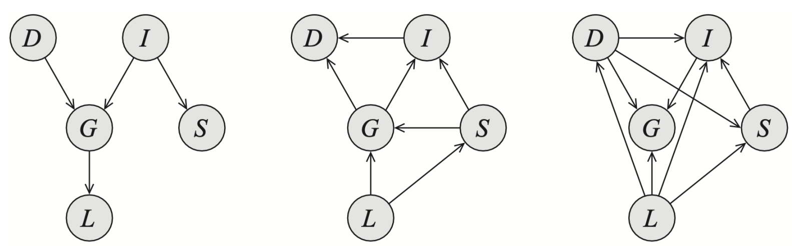
Now build the graph on $\{X_1, \dots, X_d\}$ recursively starting from isolated vertex $X_1$, then adding $X_2$ and edges $X_2 \to X_1$ if $\mathrm{Pa}_2 \neq \varnothing$, then adding $X_3$ and edges $X_3 \to X_k$ if $k \in \mathrm{Pa}_3$, etc. Clearly this graph is a DAG, factors $P$, and is minimal. $\square$

Interestingly, the constructive proof of the previous theorem immediately shows that minimal DAGs exist but not guaranteed to be unique. In fact, we have exhibited a DAG using a specific ordering of the random variables, but choosing a different ordering can lead to another minimal DAG. We illustrate this in the following example:

**Example 10.** *Taken from [KF09, Sections 3.2.1 and 3.4.1 for details]. We consider a set of random variables $\{I, D, G, S, L\}$ (Intelligence, Difficulty, Grade, SAT, Letter), and $P$ whose density is given by*

$$p(i, d, g, s, l) = p_I(i) p_D(d) p_{G|ID}(g \mid i, d) p_{S|I}(s \mid i) p_{L|G}(l \mid g) \tag{12}$$

*The following are the minimal DAGs obtained from the algorithm used in the proof of Theorem 5, using the ordering $(I, D, G, S, L)$, $(L, S, G, I, D)$, and $(L, D, S, I, G)$.*




(a) (b) (c)

**Figure 3.8 Three minimal I-maps for $P_{\mathcal{B}^{student}}$, induced by different orderings:** (a) $D, I, S, G, L$; (b) $L, S, G, I, D$; (C) $L, D, S, I, G$.

Note that, perhaps surprisingly, the three minimal DAGs $G_1, G_2, G_3$ in the previous example are not Markov equivalent (they don't have the same skeleton). There is nothing contradictory here: this is because they encode different subsets of $\mathcal{I}(G_1), \mathcal{I}(G_2), \mathcal{I}(G_3) \subset \mathcal{I}(P)$ but $\mathcal{I}(G_1) \neq \mathcal{I}(G_2) \neq \mathcal{I}(G_3)$.

This also shows that minimality and faithfulness are distinct notions, since for $P$ faithful to $G$ we must have $\mathcal{I}(G) = \mathcal{I}(P)$. Minimality is, however, a necessary condition for faithfulness.

**Proposition 1.** *If $P$ is faithful and factorizes according to $G$, then $G$ is minimal.*

*Proof.* A sktech of proof is given in [PJS17, Proposition 6.35], but it relies on the premise that two vertices with no edge between them can always be d-separated, which is still unclear to me how to prove formally. $\square$

## 2 Causal models, causes, effects, interventions

A "causal network" is a Bayesian network. The distinction between causal network and Bayesian network is purely semantic, they are mathemtically the same object. The distinction stem from the fact that in the causal network, the edges are interpreted as representing a cause-effect relation. It is indeed the DAG that defines what is a *cause* and an *effect*.

**Definition 11** (Cause)**.** *In a causal network* $(G, P)$*,* $X_i$ *is a cause of* $X_j$ *if there is a directed path from* $X_i$ *to* $X_j$*. It is a direct cause if there is an edge* $X_i \to X_j$*.*

### 2.1 "Classical" causal modeling principles

What distinguish a causal network from a Bayesian network is the meaning of the arrows. This implies that certain principles are assumed when modeling a causal networks. Those principles are not mathematical, but rather philosophical.

**Principle 1** (Reichenbach's common cause principle)**.** *If two random variables* $X$ *and* $Y$ *are statistically dependent, then either* $X$ *causes* $Y$*, or* $Y$ *causes* $X$*, or there exists a third variable* $Z$ *that causally influences both. Furthermore, this variable* $Z$ *screens* $X$ *and* $Y$ *from each other in the sense that* $X \perp Y \mid Z$*.*

**Principle 2** (Principle of independent mechanisms)**.** *The Principle of independent mechanisms posits that the mechanisms governing the generation of a system's variables are autonomous and do not influence each other. Specifically, the causal process that determines the effect of a variable* $X$ *on another variable* $Y$ *is independent of the processes governing* $X$ *itself. This principle implies that changes to one causal mechanism (e.g., intervening on* $X$*) should not alter the mechanisms governing the remaining variables.*

The principle of independent mechanisms underpins many causal inference methods by ensuring that the causal structure can be disentangled into distinct, modular components, facilitating robust predictions and transferability across different contexts.

The following is an application example of the principle of independent mechanisms.

**Example 11.** *Borrowed from [PJS17, Section 2.1]. Suppose we have estimated the joint density* $p(a, t)$ *of the altitude* $A$ *and the average annual temperature* $T$ *of a sample of cities in some country. Consider the following ways of expressing* $p(a, t)$*:*

$$p(a, t) = p(a \mid t)p(t) = p(t \mid a)p(a). \tag{13}$$

*The first decomposition describes* $T$ *and the conditional* $A \mid T$*. It corresponds to a factorization of* $p(a, t)$ *according to the graph* $T \to A$*. The second decomposition corresponds to a factorization according to* $A \to T$*. Can we decide which of the two structures is the causal one (i.e., in which case would we be able to think of the arrow as causal)?*

*A first idea is to consider the effect of interventions. Imagine we could change the altitude* $A$ *of a city by some hypothetical mechanism that raises the grounds on which the city is built. Suppose that we find that the average temperature decreases. Let us next imagine that we devise another intervention experiment. This time, we do not change the altitude, but instead we build a massive heating system around the city that raises* *the average temperature by a few degrees. Suppose we find that the altitude of the city is unaffected.*

*Intervening on $A$ has changed $T$, but intervening on $T$ has not changed $A$. We would thus reasonably prefer $A \to T$ as a description of the causal structure.*

*Why do we find this description of the effect of interventions plausible, even though the hypothetical intervention is hard or impossible to carry out in practice? If we change the altitude $A$, then we assume that the physical mechanism $p(t \mid a)$ responsible for producing an average temperature is still in place and leads to a changed $T$. This would hold true independent of the distribution from which we have sampled the cities, and thus independent of $p(a)$. Austrians may have founded their cities in locations subtly different from those of the Swiss, but the mechanism $p(t \mid a)$ would apply in both cases. If, on the other hand, we change $T$, then we have a hard time thinking of $p(a \mid t)$ as a mechanism that is still in place — we probably do not believe that such a mechanism exists in the first place. Given a set of different city distributions $p(a, t)$, while we could write them all as $p(a \mid t)p(t)$, we would find that it is impossible to explain them all using an invariant $p(a \mid t)$.*

*Our intuition can be rephrased and postulated in two ways: If $A \to T$ is the correct causal structure, then*

*(i) it is in principle possible to perform a localized intervention on $A$, in other words, to change $p(a)$ without changing $p(t \mid a)$, and*

*(ii) $p(a)$ and $p(t \mid a)$ are autonomous, modular, or invariant mechanisms or objects in the world.*

### 2.2 Alternative causal modeling principles

Occam razor's: minimum description lentgh!

### 2.3 Interventions

## 3 Causal effects

### 3.1 Nonparametric identifiability of causal effects

### 3.2 Parametric identifiability

## 4 SCM

## References

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009.

[Lau96] Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York, May 1996.

[Pea22] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge New York, NY Port Melbourne New Delhi Singapore, second edition, reprinted with corrections edition, 2022.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* Adaptive Computation and Machine Learning. The MIT press, Cambridge, Mass, 2017.