Fundação Getulio Vargas
Escola de Matemática Aplicada
Curso de Graduação em Matemática Aplicada.

Um Estudo Sobre Causalidade

Pedro Rodrigues Castilho Dall'Antonia

Rio de Janeiro - Brasil
2022

Fundação Getulio Vargas
Escola de Matemática Aplicada
Curso de Graduação em Matemática Aplicada

Um Estudo Sobre Causalidade

"Declaro ser o único autor do presente projeto de monografia que refere-se ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador."

Pedro Rodrigues Castilho Dall'Antonia

Rio de Janeiro - Brasil
2022

Fundação Getulio Vargas
Escola de Matemática Aplicada
Curso de Graduação em Matemática Aplicada

Título da dissertação

"Projeto de Monografia apresentado à Escola de Matemática Aplicada como requisito parcial para continuidade ao trabalho de monografia."

Aprovado em _____ de _____ de _____

Grau atribuído ao Projeto de Monografia: _____

Professor Orientador:
Escola de Matemática Aplicada
Fundação Getúlio Vargas

FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

PEDRO RODRIGUES CASTILHO DALL'ANTONIA

UM ESTUDO SOBRE CAUSALIDADE

Trabalho de Conclusão de Curso - TCC apresentado ao Curso de Graduação em Matemática Aplicada da Escola de Matemática Aplicada para obtenção do grau de Bacharel (a) em Matemática Aplicada.

Data da Defesa: 07/12/2022

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Luiz Max Fagundes de Carvalho
Orientador

Claudio José Struchiner
Membro

Eliezer de Souza da Silva
Membro

Dedico esse TCC aos meus pais, Glauco e Cristina, por serem a ***causa*** *de tudo que sou hoje, obrigado por todo suporte e carinho. Agradeço também à Dávila, minha namorada, por estar sempre ao meu lado, tanto nos momentos bons quanto nos ruins, dando suporte sempre que eu estava desmotivado. Finalmente, meus mais sinceros agradecimentos ao meu orientador, Dr. Luiz Max, por todos os conselhos, puxões de orelha e por deixar sempre a porta aberta de sua sala.*

Resumo

Neste trabalho, exploramos o campo da inferência causal através de modelos gráficos, desenvolvido por Judea Pearl. Começando com uma breve introdução aos conceitos de teorias dos grafos, e mergulhamos no tópico de Redes Bayesianas e seu papel em responder perguntas associativas, pertencentes ao primeiro nível da hierarquia causal.

No segundo capítulo, enriquecemos a estrutura das Redes Bayesianas com mecanismos causais, dando origem as Redes Bayesianas Causais, o que nos possibilita responder perguntas sobre intervenções a partir de dados observacionais. Por fim, visitamos o mundo dos contrafactuais e as estruturas que possibilitam seu estudo, os modelos estruturados causais. Com os resultados desenvolvidos ao longo dos capítulos, temos em mãos ferramentas próprias para o estudo da causalidade e tomada de decisão.

Notes

Sumário

1 Introdução **5**
1.1 Previsão *vs* Decisão 5
1.2 A Hierarquia de Pearl. 8

2 O Primeiro Passo Na Direção Da Causalidade **9**
2.1 Um Rápido Passeio Por Grafos 10
2.2 Redes Bayesianas 13
2.3 D-Separação 18

3 Intervenções **20**
3.1 Modelos Causais 22
3.2 Identificando perguntas causais 25
3.2.1 Voltando ao paradoxo de Simpson 26
3.2.2 Porta dos fundos 28
3.2.3 Porta da frente 30
3.3 Calculo de Do 33
3.3.1 Fórmula do ajuste revisitada 35
3.3.2 Identificabilidade 35
3.4 Intervenções Condicionais 37

4 Contrafactuais **39**
4.1 Onde Encontramos Contrafactuais? 39
4.2 Modelos Estruturais Causais 40
4.2.1 SCM no Primeiro Degrau Causal: Observações 41
4.2.2 SCM no Segundo Degrau Causal: Intervenções 42
4.2.3 SCM no Último Degrau Causal: Imaginando Contrafactuais 44

5 Conclusão **48**

1 Introdução

1.1 Previsão *vs* Decisão

Com o aumento exponencial na criação de dados globalmente, o papel do cientista de dados tem se tornado cada vez mais relevante para tomada de decisão. Atualmente, temos a nossa disposição um acervo enorme de métodos estatísticos que muitas vezes são usados como caixas pretas na esperança de achar alguma relação entre variáveis preditivas e de interesse. Com a diminuição do custo de hardware especializados, treinar modelos avançados de aprendizado de máquinas tem se tornado cada vez mais fácil, gerando previsões cada vez mais precisas sobre o mundo.

Com essa visão é possível especular que tomar decisões com base em dados se tornará agnóstico aos conhecimentos específicos de área. As vagas dos profissionais especializados estariam com seus dias contados, por que contratar algum especialista se podemos treinar modelos avançados que consideram milhares de variáveis em suas análises? Felizmente para o profissional especializado, o problema não é tão simples. Por mais que os computadores já tenham ultrapassado o poder preditivo de profissionais em diversas áreas, isso não significa que suas previsões ajudam na tomada de decisão. Essa falsa ilusão vem da hipótese de que quanto mais preciso for a previsão, mais informada será a tomada de decisão, o que não é verdade.

Tome o seguinte exemplo sobre tomada de decisão: Após o termino da graduação na FGV, uma jovem cientista de dados precisa decidir seus próximos passos, inclusive onde morar. Ela se encontra indecisa entre se mudar para São Paulo ou continuar no Rio de Janeiro. Como a jovem cientista acabou de iniciar sua carreira, ela se preocupa muito com o custo de vida e decide fazer um estudo para saber o quão caro é viver em cada cidade. Com a popularização de aplicativos de compras que monitoram seus gastos, e a facilidade de coletar os dados próprios, ela baixa todos os dados pessoais disponíveis que poderiam revelar padrões no custo de vida em todas as cidades que já visitou. Como a cientista utilizará os próprios dados, ela imagina reduzir bastante qualquer viés que teria se usasse informações publicas, já que são dados que refletem seu próprio estilo de vida e não de toda uma população.

Um dos aplicativos que ela mais usa é o de serviços de táxis. E como ela já usou o aplicativo tanto no Rio quanto em São Paulo, ela decide analisar onde é mais caro chamar uma corrida. A métrica que ela considera é reais por quilômetro de viagem, a qual ela calcula assim que termina de baixar os dados (Figura 1).

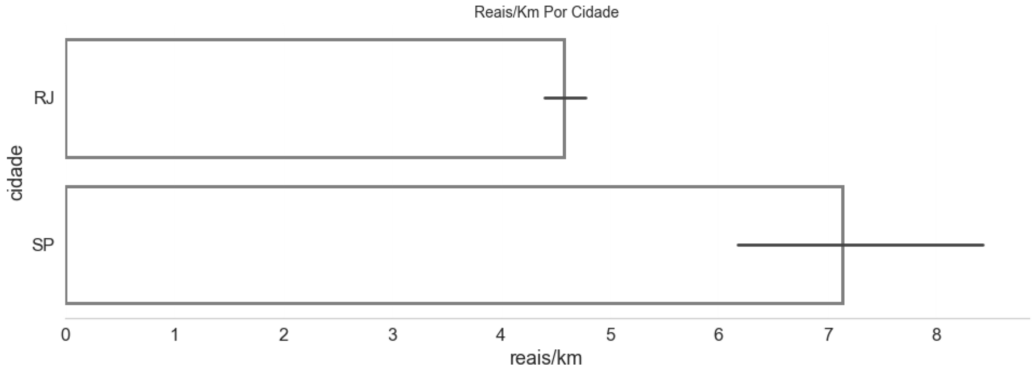

Figura 1: Gráfico acima mostra a média dos custos por quilômetro das viagens por cidade.

Sem a necessidade de nenhum modelo, já estava evidente que São Paulo era muito mais caro que Rio de Janeiro no custo de transporte de táxis pelo aplicativo. Mas uma diferença tão grande não era esperada, a média de São Paulo parece ser quase que o dobro da média do Rio nesta analise.

A cientista de dados decide montar um modelo plausível para o cálculo do custo da viagem por quilômetro percorrido para tentar entender o que pode ter dado errado. Ela modela o sistema de preços da seguinte forma: O preço da viagem é conjecturado ser dado por uma taxa fixa, mais um termo proporcional a distância percorrida.

$$\text{Preço} = C_f(\text{Cidade}) + C_d(\text{Cidade}) \cdot \text{Distância},$$

onde C_f e C_d são os custos fixos e dinâmicos que dependem de cada cidade. Para encontrar a variável de interesse dividimos dos dois lados pela distância percorrida em cada viagem e obtemos uma hipérbole, onde o custo por quilômetro se torna cada vez mais caro se a viagem for curta.

Ao explorar mais os dados, a jovem cientista encontra exatamente a relação que esperava para o custo por quilômetro das viagens. E não só isso, mas ela também constata que os valores de C_f e C_d não variam conforme as cidades, o que prova que sua análise inicial estava errado mesmo.

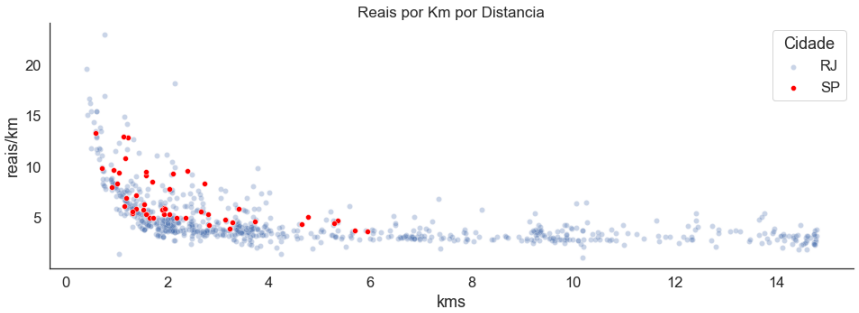

Figura 2: Grafo da relação do custo por quilômetro pela distância percorrida. Onde os pontos em azul são de viagens na cidade do Rio e pontos em vermelho viagens em São Paulo.

Mas como explicar a diferença entre as médias dos custos? Perceba como a distribuição das distâncias percorridas na (Figura 2)., são bem diferentes entre cidades, azul sendo o Rio e vermelho sendo São Paulo. Existe uma variável oculta que não estava sendo considerada e que confunde o efeito da cidade no custo, se a passageira está de férias ou não. Quando a jovem está de férias, ela prefere não sair muito do bairro que está, visitando somente lugares próximos. Mas quando ela não está de férias, ela precisava atravessar a cidade diversas vezes durante a semana, tanto para trabalhar quanto para estudar. Acontece que ela costuma visitar São Paulo com mais frequência durante suas férias, enquanto durante sua rotina normal, ela passa a maioria do tempo no Rio de Janeiro. Logo o comportamento da estudante é muito diferente em cada uma das cidades, por mais que ela conseguisse prever com precisão quanto ela gastaria na próximo viagem de férias à São Paulo em táxis, decidir para onde se mudar baseado somente em estatísticas sem entender o processo de geração dos dados não é confiável, porque seus gastos pós-decisão não são representativos dos gastos pre-decisão.

A ideia de que a distribuição das variáveis preditivas está suscetível a mudanças após uma tomada de decisão é um tema conhecido que assombra muitos modelos estatísticos.

Uma forma robusta de se proteger contra essas mudanças de distribuições é por meio de modelos causais, onde especificamos quais são as variáveis responsáveis pelas causas de outras variáveis. Ao modelar a relação funcional entre preço e distância, a jovem cientista estava especificando o processo de geração dos custos efetuados pelo aplicativo. A modelagem do processo de geração dos dados, nos responder perguntas sobre causa e efeito mesmo com acesso limitado à dados observacionais.

1.2 A Hierarquia de Pearl.

Judea Pearl, pesquisador consagrado na área de inferência causal, descobriu uma hierarquia (Figura 3) dos tipos de perguntas que podemos fazer. A medida que escalamos os degraus da hierarquia de Pearl, as perguntas trazem questionamentos cada vez mais profundos sobre a relação de causa e efeito entre variáveis.

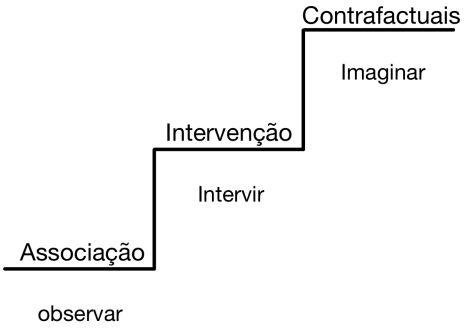

Figura 3: A hierarquia causal contem três níveis, associação, intervenção e contrafactual. A medida que escalamos na hierarquia, as perguntas dependem cada vez mais de como o processo de geração dos dados é feito.

Perguntas do primeiro nível da hierarquia, são aquelas que não questionam efeitos causais e pertencem ao degrau da associação. Essas perguntas se preocupam exclusivamente em como as variáveis estão estatisticamente associadas umas com as outras; se **observamos** uma mudança em X, o que podemos esperar ver em Y? Note que não estamos tentando entender como mudanças em X afetam Y, já que ao observar mudanças nas variáveis não estamos ativamente controlando o ambiente e outras variáveis no sistema

podem variar livremente, dificultando o estudo dos efeitos de X em Y.

O segundo nível da hierarquia é o das intervenções, onde as perguntas servem para questionar como um sistema se comporta após uma intervenção em suas variáveis. Uma intervenção é uma modificação no processo de geração dos dados de um modelo. Não estamos mais passivamente estudando um sistema, mas ativamente alterando seu comportamento, para isolar as relações de causa e efeito.

O último nível na hierarquia é um pouco mais confuso que os anteriores. Se trata de perguntas contrafactuais, perguntas sobre como o mundo seria se o passado tivesse sido diferente. Experimentos nesse nível são impossíveis, não temos como voltar ao passado e realizar experimentos para responder perguntas contrafactuais, nos resta imaginar como o mundo seria. A formalização matemática do que significa *imaginar* é deixada para o último capítulo desse estudo.

2 O Primeiro Passo Na Direção Da Causalidade

Neste capitulo falaremos sobre o primeiro elemento na hierarquia causal, o habitat dos métodos estatísticos, a associação. Muitos métodos estatísticos usam a associação entre variáveis para previsão mesmo que não haja causalidade, um exemplo clássico é estimar numero de afogamentos a partir do faturamento de uma barraca de sorvete na praia.

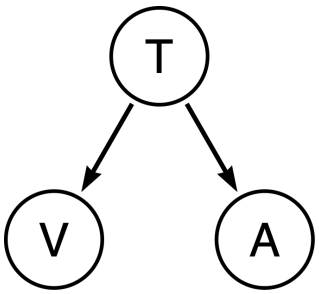

Figura 4: **Grafo que descreve a intuição que temos entre a relação de vendas de sorvetes (V), afogamentos (A) e a temperatura (T).**

Como dias quentes trazem mais pessoas para praia, mais pessoas com-

pram sorvete e mais pessoas entram no mar, aumentando o número esperado de afogamentos, mesmo não havendo uma relação causal entre às duas variáveis, vendas de sorvetes e afogamentos, elas estão correlacionadas.

Deste pequeno exemplo (Figura 4) surge uma das frases mais repetidas no começo de qualquer curso introdutório de estatística, **"correlação não implica em causalidade"**. Entretanto, por mais que este alerta seja repetido com tanta frequência, é incomum estudarmos formas de nos defender contra o falso indício de causalidade. Veremos mais tarde que a estatística sozinha é incapaz de nos proteger contra correlações espúrias e que muitas análises estão sujeitas a predições pouco robustas e até mesmo erradas.

2.1 Um Rápido Passeio Por Grafos

Para representarmos (in)dependências estatísticas e causais será útil trabalharmos com algumas definições de teoria de grafos. Em particular precisaremos definir o que é um grafo direcionado acíclico, conhecido como DAG (2.7).

Já utilizamos um DAG no nosso primeiro exemplo do capítulo e precisamos formalizar essa estrutura matemática que será base para muitas outras estruturas ao longo da nossa escalada na hierarquia causal. Começaremos definindo o que é um grafo (2.1)e um grafo direcionado (2.2).

Definição 2.1 (**Grafo**). *Um Grafo $\mathcal{G}$ é uma tupla $(\mathbf{V}, A)$ onde $\mathbf{V}$ é um conjunto onde seus elementos são chamados de nós ou vértices e A é uma relação binaria simétrica sob o conjunto $\mathbf{V} \times \mathbf{V}$. um elemento $(a, b) \in A$ é chamado de aresta de G.*

Definição 2.2 (**Grafo Direcionado**). *Um Grafo direcionado $\mathcal{G}$ é um Grafo $(\mathbf{V}, A)$ onde A é uma relação binaria anti-simétrica em $\mathbf{V} \times \mathbf{V}$. Um elemento $(a, b) \in A$ é chamado de arco, onde a é chamado de pai de b e b filho de a.*

Cada vértice do grafo pode ter múltiplos pais e filhos, e como veremos nos próximos capítulos, saber quem são todos os pais de um nó é de grande importância para isolar efeitos causais, com isso definimos (2.3). Além disso, o estudo de como informação flui entre vértices depende de caminhos entre nós no grafo, para isso definimos os conceitos de trilhas, caminhos e loops.(2.4).

Definição 2.3 (**Pais e Filhos**). *Os conjuntos:*

$$Ch(a) := \{x \mid (x, a) \in A\},$$
$$Pa(a) := \{x \mid (a, x) \in A\}.$$

São denominados, respectivamente, de filhos e pais de a.

Definição 2.4 (**Trilhas, Caminhos e Loops:**). *Uma trilha de* $\mathcal{G} = (\mathbf{V}, A)$ *é uma sequência de nós* $\{x_i\}_{i \in I}$ *onde* $(x_i, x_{i+1}) \in A$ *e cada arresta aparece no máximo uma vez na trilha. Um caminho em* $\mathcal{G}$ *é uma trilha indexada por* $I = \{1, 2...n\}$ *onde* $x_1 \neq x_n$*. Assim como um loop é uma trilha onde* $x_1 = x_n$*. Se* $\mathcal{G}$ *é um grafo direcionado, caminhos são chamados de caminhos direcionados e loops são chamados de ciclos.*

Assim como os conceitos de pais e filhos são definidos a partir da direção dos arcos de um grafo, podemos estender esses conceitos utilizando caminhos direcionados para poder falar de ancestrais e descendentes(2.5).

Definição 2.5 (**Ancestrais e Descendentes:**). *Dado um Grafo Direcionado* $\mathcal{G} = (\mathbf{V}, A)$ *e elementos* $a, b \in \mathbf{V}$*, a é descendente de b e b é ancestral de a se existe um caminho direcionado* $\{x_i\}_{i \in I}$ *onde* $x_1 = b$ *e* $x_n = a$*.*

Os conjuntos de todos os ancestrais e descendentes de a são denotados como $An(x)$ *e* $De(x)$*, respectivamente.*

Mais adiante precisaremos escolher sequências de nós para provar propriedade de modelos gráficos. Uma forma natural de escolher sequências é por meio de uma ordenação. Certo grafos possuem estrutura suficiente para ser possível ordenar seus vértices.

Definição 2.6 (**Ordem Topológica:**). *Dado um grafo* $\mathcal{G} = (\mathbf{V}, A)$*, uma sequência de vértices* $\{x_1, x_2 \ldots, x_n\}$ *é dita uma ordenação topológica se* $(x_i, x_j) \in A$ *implica que* x_i *vem antes de* x_j *na sequência.*

A ordem topológica é coerente com os caminhos direcionado de um grafo, mas não necessariamente existe uma única ordem, um grafo pode induzir diversas ordens topológicas distintas, como no exemplo abaixo (5):

Finalizamos as definições com o conceito de um dos objetos centrais deste estudo, o DAG (2.7).

Definição 2.7 (**Grafo direcionado acíclico (DAG)**). *Um grafo* $\mathcal{G} = (\mathbf{V}, A)$ *direcionado é uma DAG se e somente se* $\forall x \in \mathbf{V}, x \notin An(x)$*.*

Uma propriedade importante dos DAGs é que sempre conseguiremos achar uma ordem topológica valida.

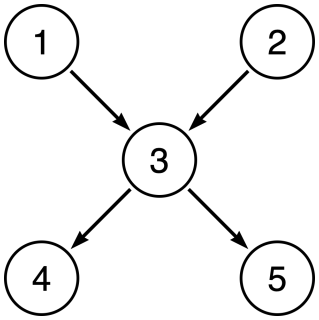

Figura 5: No grafo acima tanto a sequencia {1,2,3,4,5} quanto {2,1,3,5,4} são ordens topológicas validas.

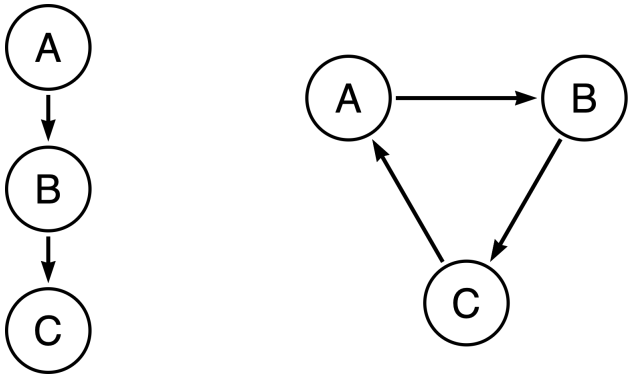

Figura 6: O grafo da direita é um DAG. Já o da esquerda não é, porque possui um ciclo.

Teorema 2.1. *Toda DAG possui ao menos uma ordenação topológica.*

A prova desse teorema se baseia em um simples algoritmo [8, p. 1146] que encontra ordens topológicas, se o algoritmo terminar de rodar, quer dizer que ele encontrou uma ordenação valida. A ideia do programa é criar uma lista L com todos os vértices e outra lista ordenada vazia O, escolhemos um vértice que não possui pais em L e o retiramos de L e adicionamos como o último elemento de O. Repetimos esse processo até L ficar vazia.

Ao longo dessa revisão, sempre modelaremos nossos problemas com um DAG, onde vértices representam variáveis aleatórias, como temperatura e faturamento na figura 4 e seus arcos, a existência de alguma relação entre as

variáveis[1]. A partir dos DAGs definiremos Redes Bayesianas, Redes Bayesianas Causais e SCM, cada uma delas sendo a estrutura que permitirá o estudo em cada um dos degraus da hierarquia causal.

2.2 Redes Bayesianas

Dado um modelo com N variáveis aleatórias e sua distribuição de probabilidade conjunta $P(X_1, ...X_n)$, podemos responder qualquer pergunta do primeiro nível da hierarquia de Pearl[2], perguntas da forma "Qual a probabilidade de observar X dado que observamos Y". Mas mesmo nos modelos mais simples especificar a distribuição de probabilidade pode se tornar computacionalmente impossível [8, p. 45]. Tome o simples caso de N variáveis binarias $X_1, ...X_n$, o total de configurações existentes no modelo são 2^n, ou seja, para especificar a distribuição deste modelo precisamos armazenar $2^n - 1$ valores[3]. Além disso, estimar a distribuição de probabilidade demandaria uma quantidade ainda maior de dados [8, p. 45].

Como de costume em muitas áreas da matemática, quando nos deparamos com um problema muito complicado, adicionamos hipóteses para simplificar o problema. Neste caso participar as hipóteses adicionais serão sobre independências condicionais entre variáveis, da forma $\{X \perp\!\!\!\perp Y \mid Z\}$, as quais em muitos casos são razoáveis de serem incluidas no modelo. As hipóteses de independência condicional ajudaram a reduzir a dimensão do espaço onde os parâmetros da distribuição de probabilidade do modelo se encontram. Tome novamente um modelo de N variáveis binarias $X_1, ...X_n$ com distribuição conjunta P, mas dessa vez representando lançamentos de moedas. Toda distribuição $P(X_1, ..., X_n)$ pode ser fatorada como:

$$P(X_1, X_2, ...X_n) = P(X_1)P(X_2 \mid X_1)...P(X_n \mid X_1, ...X_{n-1}).$$

Para especificar cada um dos termos $P(X_j \mid X_1, ...X_{j-1})$ precisamos de um parâmetro para cada possível estado de $(X_1, ...X_{j-1})$, ou seja, 2^{j-1}

[1]Definiremos que tipo de relação é essa no futuro, ela pode ser tanto estatística quanto causal dependendo do modelo que estamos utilizando.

[2]Como Judea Pearl foi o primeiro a apresentar e discutir hierarquia causal, é comum atribuir a descoberta a ele. [2]

[3]Precisamos definir a probabilidade de observar todos os possíveis estados A_i do modelo com exceção de uma configuração, pois após encontrar a probabilidade de $2^n - 1$ estados, a probabilidade do último estado precisa ser $1 - \sum_{i=1}^{2^n-1} P(A_i)$ para que a soma das probabilidades de todas configurações seja 1.

parâmetros para especificar esse fator. Desta forma precisamos de um total de $\sum_{i=0}^{n-1} 2^{i-1} = 2^n - 1$ valores, como já tínhamos calculados, para especificar toda a distribuição conjunta. Como estamos lidando com lançamentos de moedas, é natural supor que $X_j \perp\!\!\!\perp X_i, \forall i \neq j$, isso simplifica:

$$P(X_j \mid X_1, ...X_{j-1}) = P(X_j).$$

Desta forma reduzimos a necessidade de 2^{j-1} parâmetros para somente $2^0 = 1$ por função de distribuição condicional. Agrupando todos os termos, nosso modelo pode ser totalmente especificado com somente n parâmetros.

Note como a redução na complexidade do modelo vem da forma como escrevemos sua fatorização:

$$P(X_1, X_2, ...X_n) = P(X_1)P(X_2 \mid X_1)...P(X_n \mid X_1, ...X_{n-1}).$$

$$P(X_1, X_2, ...X_n) = P(X_1)P(X_2)...P(X_n).$$

E como cada fatorização induz um grafo direcionado acíclico (Figura 7), onde as funções de distribuição condicional das variáveis dependem dos pais no grafo.

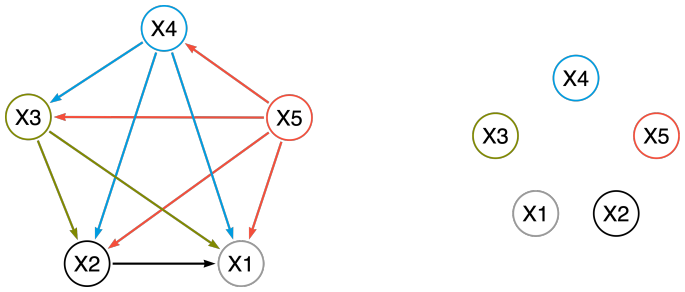

Figura 7: **Grafo induzido da fatorização da probabilidade conjunta de um modelo**. As cores mostram as dependências das funções de probabilidade condicional de cada variável.

O surgimento natural dos grafos entre expressões de distribuições de probabilidade condicional nos motiva a definir a estrutura central deste capitulo.

Definição 2.8 (**Fatorização Induzida Por Grafos**). *Dado um grafo $\mathcal{G}$ onde os vértices $X_1, X_2..., X_n$ são variáveis aleatórias com distribuição P,*

dizemos que P fatora de acordo com $\mathcal{G}$ se P pode ser expressa como:

$$P(X_1, X_2..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)).$$

Onde $Pa(X_i)$ é o conjunto de variáveis que são pais de X_i no grafo $\mathcal{G}$.

Definição 2.9 (**Rede Bayesiana**). *Uma Rede Bayesiana é uma tupla $\mathcal{B} = (\mathcal{G}, P)$ onde $\mathcal{G}$ é uma DAG com variáveis aleatórias $X_1, X_2..., X_n$ como vértices, P fatora de acordo com $\mathcal{G}$ e P é especificada por um conjunto de funções de probabilidade condicional.*

Para ilustrar algumas propriedades de Redes Bayesianas, apresento o seguinte modelo tirado do livro de Koller e Friedman [8, p. 48]. Uma empresa decide abrir novas vagas de emprego e seu objetivo é contratar os alunos mais inteligentes, mas como inteligencia (I) não é mensurável diretamente precisamos de um modelo que possibilite estimá-la indiretamente por meio de variáveis que são observáveis. A empresa tem acesso a quatro variáveis:

- Posição no ranking no vestibular (V);
- Boletim escolar (B);
- Dificuldade do curso (D);
- Carta de recomendação (C).

E a empresa decide utilizar o seguinte grafo $\mathcal{G}_\mathcal{B}$ (Figura 8) para modelar o problema:

Onde cada arresta representa uma função de distribuição de probabilidade condicional conhecida[4] de tal forma que $P(D, I, B, V, C)$ fatora de acordo com $\mathcal{G}_\mathcal{B}$. O objetivo da empresa é calcular $P(I \mid C, B, D, V)$ para cada candidato de forma que suas escolhas sejam as mais embasadas possíveis. Mas coletar, tratar e armazenar todos esses dados pode ser muito custoso, será que é possível descartar informação de forma que $P(I \mid C, B, D, V)$ ainda seja calculável?

Intuitivamente podemos desconfiar que se já tivermos informação sobre o boletim do aluno, a carta de recomendação não carregaria informação sobre

[4]As distribuições de probabilidade podem ser estimadas por meio de fontes externas ou com ajuda especialistas

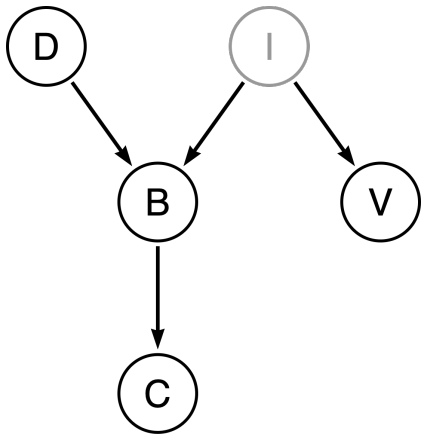

Figura 8: **O grafo representa as hipóteses de independências condicionais do modelo**. Os nós em destacados representam as variáveis que são observáveis.

o quão inteligente o aluno é, já que o professor se baseia no boletim para escrever a carta. E de fato, isso é verdade e provamos abaixo:

$$
\begin{aligned}
P(C \mid B, I, D) &= \frac{P(C, B, I, D)}{P(B, I, D)}, \\
&= \frac{\sum_r P(C, B, I, D, r)}{\sum_{r,c} P(B, I, D, c, r)}, \\
&= \frac{\sum_r P(D)P(I)P(B \mid D, I)P(C \mid B)P(r \mid I)}{\sum_{r,c} P(D)P(I)P(B \mid D, I)P(c \mid B)P(r \mid I)}, \\
&= \frac{P(D)P(I)P(B \mid D, I)P(C \mid B)\sum_r P(r \mid I)}{P(D)P(I)P(B \mid D, I)\sum_{r,c} P(c \mid B)P(r \mid I)}, \\
&= \frac{P(D)P(I)P(B \mid D, I)P(C \mid B)}{P(D)P(I)P(B \mid D, I)}, \\
&= P(C \mid B).
\end{aligned}
$$

Dai segue que[5]:

[5]Omitimos as propriedades de independência condicional, mas elas podem ser encontradas em [8, p. 25]

$$P(C \mid B, I, D) = P(C \mid B) \implies C \perp\!\!\!\perp \{I, D\} \mid B \implies C \perp\!\!\!\perp I \mid B \implies I \perp\!\!\!\perp C \mid B.$$

Além de mostrar que $C \perp\!\!\!\perp I \mid B$, a prova também mostra que $C \perp\!\!\!\perp \{I, D\} \mid B$, e ainda conseguimos mostrar que C é independente de todas as outras variáveis dado B. Esse fato é consequência de um teorema ainda mais geral sobre redes bayesianas, se fixarmos os valores dos pais de uma vértice X então essa variável é independente de todos os vértices que não são seus descendentes:

Teorema 2.2 (**Condição de Markov Local**). *Dado uma Rede Bayesiana $\mathcal{B}$ que fatora de acordo com $\mathcal{G}$ e possui vértices $\mathbf{V} = \{X_1, ...X_n\}$ é sempre verdade que:*

$$X_i \perp\!\!\!\perp \mathbf{V} - De(X_i) \mid Pa(X_i)$$

Ou seja, X_i é independente de seus não descendentes condicionados aos seus pais. O conjunto de todas as proposições desta forma é conhecido como independências locais do grafo sendo denotado por $\mathcal{I}_l(\mathcal{G})$.

Esse teorema nos mostra que a partir do grafo de uma rede bayesiana, independente de qual seja a distribuição conjunta, sabemos que as independências locais são sempre validas. Chamamos essa propriedade de I-map e definimos formalmente como:

Definição 2.10 (I-map). *Dado um grafo $\mathcal{G}$ e uma distribuição P, denotamos $\mathcal{I}(\mathcal{G})$ e $\mathcal{I}(P)$ como os conjuntos de afirmações verdadeiras sobre as independências induzidas pelo grafo G e contidas em P, respectivamente. Dizemos que $\mathcal{G}$ é um I-map de P se $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(P)$*

O Teorema 2.2 mostra que se P fatoriza de acordo com $\mathcal{G}$ então $\mathcal{G}$ é um I-map de P. Mas a volta também é valida:

Teorema 2.3 (I-map $\implies$ Fatorização). *Dado um grafo $\mathcal{G}$ e uma distribuição P, se $\mathcal{G}$ for I-map de P então P fatoriza de acordo com $\mathcal{G}$.*

A prova do teorema acima pode ser encontrada no livro do Koller e Friedman [8, p. 62] O teorema 2.2 e 2.3 nos da uma nova forma de caracterizar Redes Bayesianas, podemos tanto defini-las a partir de sua fatorizações quanto a partir de suas hipóteses de independências locais.

Uma pergunta natural que surge é se conseguimos representar todas as independências de P dentro do grafo. Para conseguir falar de todas as independências induzidas por um grafo $\mathcal{G}$ precisamos definir um novo conceito, d-separação.

2.3 D-Separação

Existem três estruturas básicas chaves para entender independências condicionais em grafos:

1. Correntes: $A \rightarrow B \rightarrow C$.
2. Causa Comum: $A \leftarrow B \rightarrow C$.
3. Efeito Comum: $A \rightarrow B \leftarrow C$.

As duas primeiras estruturas se comportam de forma muito parecidas. Como vimos, no exemplo da rede do estudante, se condicionarmos em B, A e C se tornam independentes um do outro. Estamos basicamente bloqueando o "fluxo" de informação entre A e C. Mas o oposto acontece em estruturas de efeito comum, assim que observamos B; A e C se tornam dependentes.

Informalmente, quando informação consegue "fluir" entre os nós A e C por B dizemos que a trilha $A - B - C$ está ativa, sendo assim, se A e C forem independentes dado B, dizemos que a trilha está bloqueada. A mesma noção é valida para trilhas com mais de 3 nós, para que uma trilha $X_1 - ... - X_n$ esteja ativa, intuitivamente[6], é preciso que não exista nenhuma trilha $X_i - X_{i+1} - X_{i+2}$ bloqueada [8, p.71]. Com essa ideia de caminhos bloqueados e ativos introduzimos a definição de d-separação:

Definição 2.11 (D-Separação). *Um caminho P é bloqueado por um conjunto de nós Z se e somente se:*

1. *P contem alguma estrutura de corrente $A \rightarrow B \rightarrow C$ ou causa comum $A \leftarrow B \rightarrow C$ onde $B \in Z$.*
2. *P contem uma estrutura de efeito comum tal que $B \notin Z$ e nenhum descendente de B pertence à Z.*

Se Z bloqueia todos os caminhos entre dois nós X e Y, então dizemos que X e Y estão D-separados dado Z.

Na estrutura abaixo (Figura 9), por exemplo, X e Y estão D separados, pois existe um efeito comum $X \rightarrow W \leftarrow Z$, pois ela bloqueia todos os caminhos entre X e Y, é fácil ver, pois só existe um caminho entre eles.

[6] D-separação serve para formalizar essa intuição

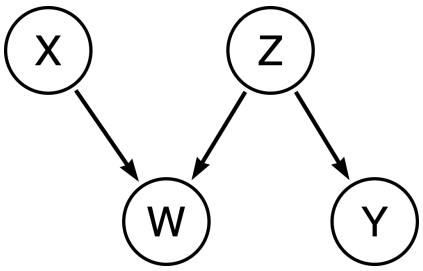

Figura 9: X e Y estão d-separados, mas deixam de ficar se condicionarmos em W, ficam novamente se condicionar em Z.

Agorá X e Y não estão d-separados quando condicionamos em W, justamente porque W é um efeito comum. Mas se condicionarmos novamente em W e Z, X e Y se tornam d-separados mais uma vez.

Um teorema importante da d-separação é que se dois vértices estiverem d-separados dado Z, então eles são condicionalmente independentes dado Z [8, p. 72].

Teorema 2.4 (**D-Separação $\implies$ Independeria**). *Se duas variáveis X, Y estiverem D-separadas em $\mathcal{G}$ dado Z, então elas são condicionalmente independentes dado Z.*

Infelizmente o contrário não é verdade (Figura 10). Uma distribuição pode fatorar de acordo com um grafo $\mathcal{G}$ e ter duas variáveis aleatórias independentes condicionalmente, mas não d-separadas, tome o seguinte exemplo, onde os nós Z e W tomam os valores $X + U$ e $-X + U$, respectivamente, e $Y = Z + W = 2U$, onde U é um ruído qualquer. X e Y possui dois caminhos ativos entre eles, logo não estão d-separadas, mesmo assim X e Y são independentes um do outro.

Um detalhe muito importante nas estruturas de correntes é que a fatorização independe da direção dos arcos. Como a única suposição de independência condicional nas correntes é a de que $A \perp\!\!\!\perp C \mid B$, isso quer dizer que a probabilidade conjunta de $A \to B \to C$ e $A \leftarrow B \leftarrow C$ podem ambas serem fatoradas como:

$$P(A, B, C) = P(A)P(B \mid A)P(C \mid B).$$

Sendo assim, com Redes Bayesianas ficamos restritos a perguntas associativas e precisaremos redefinir o significado dos arcos para podermos responder perguntas causais.

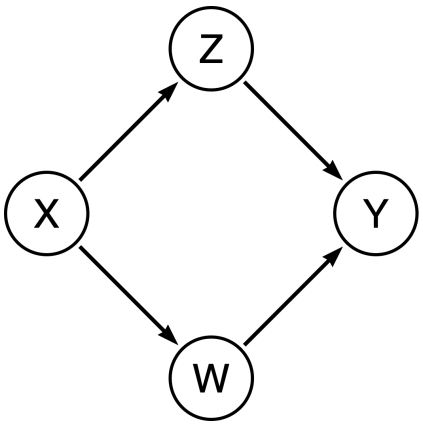

Figura 10: Mesmo não estando d-separados, as variáveis X e Y são independentes

3 Intervenções

Neste capitulo buscaremos algo mais profundo que a associação, a causa! Em diferentes áreas do conhecimento a necessidade de entender causa e efeito é fundamental. Em Medicina pode ser o efeito que administrar um medicamento terá no paciente, em Economia, o efeito do aumento dos juros na inflação, no geral é entender as consequências de nossos atos.

É fácil ver que o estudo da causalidade é de muita importância em diversas áreas, por isso nesse capitulo daremos atenção a como identificar efeitos causais, respondendo perguntas como:

"Em quanto fumar aumenta a probabilidade de um indivíduo desenvolver câncer?"

A forma tradicional de identificar e estimar efeitos causais é por meio de estudos randomizados controlados. Que se a base na ideia que se formarmos dois grupos de indivíduos grandes o suficiente a ponto que sejam estatisticamente indistinguíveis, podemos medir o efeito de um tratamento expondo um dos grupos a ele e observando as diferenças geradas. Como os dois grupos eram equivalentes no começo do teste, as diferenças entre os grupos, pós-teste, deve ser por conta do tratamento.

No entanto, muitas vezes, realizar testes randomizados é simplesmente inviável. Testar se renda universal diminuiria o analfabetismo ao longo de

gerações, seria um teste muito caro. Testar se tratamentos de choque são mais eficazes que técnicas comuns seria anti-ético de praticar, e verificar se governos ditatórias diminuem a criminalidade é simplesmente impossível de colocar a prova no formato de um teste randomizado.

Em situações onde não podemos realizar experimentos, testar causalidade se torna uma tarefa muito perigosa se não estivermos armados com uma teoria que possibilite um *framework* de identificação. Muitas teorias causais foram desenvolvidas ao longo dos anos, uma das mais conhecidas e usadas por economistas e cientistas sociais e o modelo causal de Rubin, ou *potential outcomes* (PO) [7]. Neste trabalho focaremos nos modelos gráficos de Pearl [10], também conhecido como *Structural Causal Models* (SCM) como *framework* causal. Cada um dos *frameworks* possui suas vantagens e desvantagens, SCM são muito expressivas e permitem que nossas hipóteses estejam bem claras, no entanto, as duas teorias são equivalentes, um teorema provado em SCM pode ser provado em PO e o contrário também é verdade [10, cap. 7.7.4].

Antes de introduzirmos o estudo de intervenções em redes causais, mostraremos como a falta de uma base teórica pode nos enganar ao tentar tomar uma decisão.

Exemplo 3.1 (**Paradoxo de Simpson** [12])**.** *Considere seguinte o problema, um hospital precisa escolher qual medicamento aplicar para melhor combater uma doença, a escolha deve ser feita entre a droga A e B. Neste estudo não foi conduzido testes randomizados, então o hospital tem somente informação recolhida pela observação de médicos. Essas informações incluem se o paciente se recuperou da doença (R), se ele era idoso (I), e qual droga foi administrada (D).*

Para ajudar na escolha do medicamento, o hospital contrata dois estatísticos, o primeiro mostra que os pacientes que utilizaram a droga A tiveram uma probabilidade maior de se recuperarem comparados aos pacientes que utilizaram a droga B:

$$P(R = 1 \mid D = A) > P(R = 1 \mid D = B).$$

No entanto, o segundo estatístico mostra que analisando separadamente por idade as probabilidades se invertem:

$$P(R = 1 \mid D = A, I = 1) < P(R = 1 \mid D = B, I = 1).$$

$$P(R = 1 \mid D = A, I = 0) < P(R = 1 \mid D = B, I = 0).$$

No primeiro momento o hospital assume que deve ter algum erro na análise de um dos estatísticos, como que um medicamento pode ser o melhor para toda a população, mas, simultaneamente, ter desempenho pior em todos os grupos analisados?

O caso acima é uma manifestação de um famoso fenômeno conhecido como paradoxo de Simpson[7]*. O paradoxo de Simpson não é um paradoxo estatístico como muitos acreditam, mas sim uma consequência da falta de modelagem causal. A solução do paradoxo não pode ser encontrada nos dados e sim na geração deles [5]. Voltaremos à solução do paradoxo de Simpson nas próximas seções.*

3.1 Modelos Causais

Antes de definir modelos causais precisamos definir o que significa dizer que X causa Y.

Definição 3.1. *Na definição intervencionista de causalidade, X é uma causa de Y se observamos uma variação em Y ao variamos X, mantendo todas as outras variáveis fixas.*

No exemplo do capítulo passado mostramos ser possível estimar o número de afogamentos observando as vendas de sorvetes por existir um caminho não bloqueado entre as variáveis vendas e afogamentos pelo caminho $V \leftarrow T \rightarrow A$.

Mas mesmo existindo correlação entre as variáveis, intuitivamente não existe causalidade entre elas, não seria razoável assumir que se os vendedores fizessem uma promoção de sorvetes, levando os banhistas a consumirem mais sorvetes, os bombeiros ficariam preocupados que o número de afogamentos aumentaria. Da mesma forma que para aumentar as vendas, sorveterias não cogitam afogar pessoas. A ideia de realizar uma promoção de sorvetes é equivalente a retirar o arco entre V e seu pai T, destruindo a relação estatística que existia entre V e A.

A ideia de retirar a influência dos pais de um nó em seus filhos é conhecida como uma intervenção e para representar uma intervenção utilizamos o operador de *do*. No exemplo, a intervenção em $\mathcal{G}$ é da forma $do(V = v)$, onde fixamos o valor de $V = v$, que seria o valor promocional do sorvete, e

[7] Mais sobre o paradoxo de Simpson pode ser encontrado em [13]

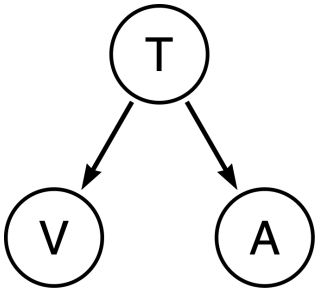

Figura 11: **Grafo $\mathcal{G}$ que descreve a relação entre vendas de sorvetes (V), afogamentos (A) e temperatura (T)**. A estrutura em V do modelo induz uma associação espuiria entre as variáveis V e A, quando observamos mudanças em V esperamos ver mudanças em A, mesmo que não existe uma relação de causa e efeitos entre as variáveis.

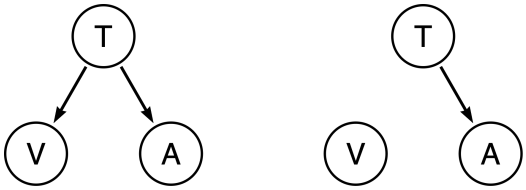

Figura 12: A esquerda temos o grafo usual $\mathcal{G}$ que modela as dependências entre as variáveis e a direita $\mathcal{G}_{V=v}$ o grafo mutilado, quando destruímos a dependência entre a venda de sorvetes e a temperatura por meio de uma intervenção.

o resultado disso é uma modificação de $\mathcal{G}$ em $\mathcal{G}_{V=v}$. Uma propriedade importante que desejamos nas intervenções é que elas sejam locais, eliminando somente o mecanismo causal entre a variável afetada e seus pais, sem alterar a probabilidade condicional entre quaisquer outros nós. Encapsulamos essa motivação na seguinte definição:

Definição 3.2 (**Rede Bayesiana Causal [RBC]**). *Seja $\mathcal{B} = (\mathcal{G}, P)$ uma rede bayesiana e $P_{X=x}(v)$ a distribuição resultante da intervenção $do(X = x)$ que fixa os valores de um subconjunto $X \subset G$ em x.*

Denotamos $\mathbf{P}_*$ o conjunto de todas as distribuições que fatoram de acordo com $\mathcal{G}_{X=x}$ para qualquer valor de x, incluindo $P(v)$ a distribuição original de $\mathcal{G}$ que corresponde a intervenção $do(X = \emptyset)$.

$\mathcal{B}$ é dita uma rede bayesiana causal de acordo com $\mathbf{P}_*$ se e somente se as três condições abaixo forem respeitadas para todo $P \in \mathbf{P}_*$:

1. *$P_{X=x}(v)$ fatora de acordo com $\mathcal{G}$*
2. *$P_{X=x}(v_i) = 1$, $\forall V_i \in X$ quando v_i for consistente com a intervenção, ou seja, $X = x$.*
3. *$P_{X=x}(v_i \mid Pa(V_i)) = P(v_i \mid Pa(V_i)), \forall V_i \notin X$ quando $Pa(V_i)$ for consistente com a intervenção.*

Para deixar a notação mais clara definimos a probabilidade:

$$P(v \mid do(X = x)) := P_{X=x}(v).$$

Dessa forma fica mais claro distinguir perguntas causais de associativas como:

"Qual a probabilidade de que haja poucos afogamentos na praia dado que poucos sorvetes foram vendidos?"

"Qual a probabilidade de que haja poucos afogamentos na praia se proibimos a venda de sorvetes?"

A primeira pergunta pode ser respondida com o cálculo da probabilidade $P(A = \text{poucos} \mid V = \text{poucos})$ no grafo $\mathcal{G}$. Já a segunda pergunta é respondida de forma bem diferente, em particular com $P(A = \text{poucos} \mid do(V = \text{poucos}))$, que representa a probabilidade condicional no grafo mutilado $\mathcal{G}_{V=v}$.

Um detalhe importante de se notar é que a intervenção $do(X = x)$ é uma idealização; muitas intervenções não são fisicamente possíveis de serem realizadas. Tome o seguinte exemplo: queremos saber o efeito que um gene específico tem no crescimento de cabelo, a intervenção correspondente seria alterar o DNA de cada célula de um ser humano. Ou qual o efeito de elevar a altitude da cidade do rio em 3000 metros no seu clima. A viabilidade das intervenções não é um obstaculo para seu estudo, até mesmo porque o cálculo das probabilidades dada por uma intervenção sempre serão feitas a partir de probabilidades condicionais correspondentes ao grafo original onde as observações foram feitas.

3.2 Identificando perguntas causais

Já temos a definição de uma RBC 3.2 e uma notação coerente que nos auxilia a formulação correta de perguntas causais, nos resta saber como calcular as probabilidades do formato $P(X = x \mid do(Y = y), Z = z)$. Como muitas vezes não podemos amostrar do grafo mutilado $G_{Y=y}$, precisamos formular perguntas causais em função de perguntas associativas, esse é conhecido como o problema de identificação e veremos quando e como podemos resolvê-lo.

A definição 3.2 foi útil por que incapsula como queremos que intervenções se comportem, no entanto, ela não é amigável para provar propriedades. Felizmente, assim como Redes Bayesianas, RBCs podem ser caracterizadas por sua fatorização [10, p. 24].

Teorema 3.1 (**Fatorização Truncada**). *$\mathcal{B} = (\mathcal{G}, P)$ é uma RBC se e somente se podemos fatorar P como:*

$$P_{X=x}(v) = \prod_{i|V_i \notin X} P(v_i \mid Pa(V_i)) \cdot \delta_x(X)$$

Com a fatorização truncada derivamos duas propriedades esperadas de RBCs [10, p. 24]:

1. $\forall V \in \mathcal{G}$:
   $P(V \mid do(Pa(V))) = P(V \mid Pa(V))$
2. $\forall V \in \mathcal{G}, \forall S \subset \mathcal{G}$ tal que $S \cap \{V, Pa(V)\} = \emptyset$:
   $P(V \mid do(Pa(V), S)) = P(V \mid do(Pa(V)))$.

A primeira propriedade nos garante que os mecanismos causais entre a variável que intervimos e seus filhos não mudam. Ou seja, modelo é agnóstico quanto a como suas variáveis tomaram um valor que tomaram, se foi por motivos naturais ou artificiais, se esse não fosse o caso, não seria possível responder perguntas sobre intervenções somente com dados observacionais, pois seriam sistemas com comportamentos diferentes. A segunda propriedade expressa a noção de invariância, a partir do momento que controlarmos os valores de $Pa(V)$, qualquer outra intervenção não afetará a probabilidade de V, porque tudo que controla o comportamento de V já foi fixado, a menos de variáveis latentes.

3.2.1 Voltando ao paradoxo de Simpson

No começo do capítulo 3.1 nos deparamos com um problema de decisão, dado um estudo observacional de duas drogas, A e B, qual deveríamos escolher como padrão nos tratamentos. No começo do capítulo não sabíamos escolher entre as duas drogas, já que a eficácia relativa entre as drogas dependia se condicionávamos por idade dos pacientes ou não. Agora que temos uma linguagem clara para formular perguntas causais, sabemos que estamos interessados em calcular $P(R = 1|do(D = A)) - P(R = 1 \mid do(D = B))$, a diferença entre a probabilidade de um indivíduo se recuperar se tratado com a droga A e com a droga B. Para encontrar o valor dessa probabilidade dessa intervenção utilizaremos a fórmula do ajuste.

Teorema 3.2 (**Fórmula do ajuste**). *Dada uma RBC $\mathcal{B}$ a fórmula do ajuste afirma que:*

$$P(Y = y \mid do(X = x)) = \sum_z \frac{P(X = x, Y = y, Pa(X) = z)}{P(X = x \mid Pa(X) = z)}.$$

Prova 3.1. *Pelo Teorema da fatorização truncada 3.1 sabemos que:*

$$P(v_1, v_2, ..., v_n \mid do(X = x)) = \prod_{i|V_i \notin X} P(v_i \mid Pa(V_i)) \cdot \delta_x(X).$$

A partir dai segue que:

$$\begin{aligned}
P(v_1, v_2, ..., v_n \mid do(X = x)) &= \prod_{i|V_i \notin X} P(v_i \mid Pa(V_i)) \cdot \delta_x(X), \\
&= \delta_x(X) \cdot \prod_{i|V_i \notin X} P(v_i \mid Pa(V_i)), \\
&= \delta_x(X) \cdot \frac{\prod_{i|V_i \notin X} P(v_i \mid Pa(V_i)) \prod_{i|V_i \in X} P(v_i \mid Pa(V_i))}{\prod_{i|V_i \in X} P(v_i \mid Pa(V_i))}, \\
&= \delta_x(X) \cdot \frac{\prod_i P(v_i \mid Pa(V_i))}{\prod_{i|V_i \in X} P(v_i \mid Pa(V_i))}, \\
&= \frac{P(v_1, ..., v_n, X = x)}{P(X = x \mid Pa(X))}.
\end{aligned}$$

Assim temos que:

$$
\begin{aligned}
P(Y = y \mid do(X = x)) &= \sum_z P(Y = y, Pa(X) = z \mid do(X = x),) \\
&= \sum_z \frac{P(Y = y, X = x, Pa(X) = z)}{P(X = x \mid Pa(X) = z)}.
\end{aligned}
$$

□

Com a fórmula do ajuste 3.2 voltamos ao problema de decisão do hospital. Podemos modelar o problema do hospital 3.1 com a seguinte RBC:

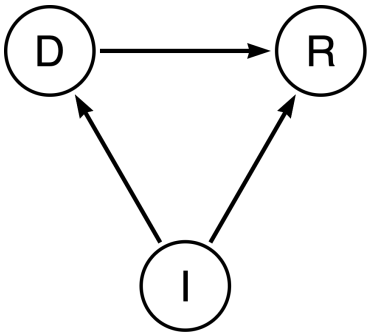

Figura 13: O Grafo do modelo de tratamento de uma doença, com as variáveis: droga utilizada no tratamento (D), se o paciente se recuperou ou não (R) e se o paciente é idoso (I). Os arcos indicam como as variáveis estão causalmente relacionadas.

Com a fórmula do ajuste temos que:

$$
\begin{aligned}
P(R = 1 \mid do(D)) &= \sum_i \frac{P(R = 1, D, I = i)}{P(D = A \mid I = i)}, \\
&= \sum_i P(I = i) \cdot P(R = 1 \mid D, I = i).
\end{aligned}
$$

Logo percebemos que para esse cenário particular o correto era comparar o resultado do medicamento condicionando a idade dos pacientes, sendo assim a droga B seria a melhor opção.

3.2.2 Porta dos fundos

A preocupação central do estudo da identificação é conseguir separar o efeito associativo do efeito causal entre duas variáveis X e Y. A fórmula do ajuste aproveita do fato de que, quando não condicionamos em nenhuma outra variável, toda associação espúria deve percorrer primeiro os arcos que ligam as variáveis do tratamento, X, aos seus pais, $Pa(X)$, até chegar em Y, nossa variável de interesse. Logo, condicionando no conjunto $Pa(X)$, estamos efetivamente bloqueando todos os caminhos do fundo de X até Y.

Infelizmente a fórmula do ajuste requer que todo o conjunto $Pa(X)$ seja observável para podermos condicionar nele, o que muitas vezes simplesmente não é possível, tornando a fórmula do ajuste inacessível. Uma ideia para contornar o problema é encontrar outro subconjunto de nós que faça o papel dos pais da variável que desejamos intervir, bloqueando novamente os caminhos do fundo entre o tratamento e resposta, Um possível conjunto de nós que concretize isso são aqueles que satisfazem o critério de porta dos fundos.

Definição 3.3 (**Critério de porta dos fundos**). *Dado uma DAG $\mathcal{G}$ e um par ordenado (X, Y) de nós pertencentes à $\mathcal{G}$, um subconjunto Z de nós de $\mathcal{G}$ é chamado de porta dos fundos de (X, Y) se respeitar todas as seguintes condições:*

1. *Nenhum nó em Z é descendente de X.*
2. *Z bloqueia todos os caminhos entre X e Y que contenham arcos direcionados à X.*

A primeira condição do critério da porta dos fundos garante que não estamos bloqueando um caminho direto entre X e Y. A segunda condição garante que associações espúrias não estão afetando Y. Com essas condições podemos enunciar outra fórmula do ajuste.

Teorema 3.3 (**Ajuste pela porta dos fundos**). *Dado uma DAG $\mathcal{G}$ e um par ordenado (X, Y) de nós pertencentes à $\mathcal{G}$, se um subconjunto Z de vértices $\mathcal{G}$ for uma porta dos fundos de (X, Y) temos que:*

$$P(Y \mid do(X = x)) = \sum_z P(z)P(Y \mid x, z).$$

Prova 3.2. *Sabemos pela fórmula do ajuste 3.2 que:*

$$\begin{aligned}
P(y \mid do(x)) &= \sum_t P(t)P(y \mid x, t), \\
&= \sum_t P(t) \sum_z P(y, z \mid x, t), \\
&= \sum_t P(t) \sum_z P(z \mid x, t)P(y \mid x, t, z).
\end{aligned}$$

Onde $T = Pa(X)$. Como Z obedece aos critérios de porta dos fundos, seus nós não são descendentes de X, logo é direto que $\{X \perp\!\!\!\perp Z \mid T\}$, podemos então simplificar $P(z \mid x, t)$ por $P(z \mid t)$. Além disso, como Z bloqueia todos os caminhos que partem de T até Y sem passar por X e X bloqueia todos os caminhos restantes até Y, temos que $\{T \perp\!\!\!\perp Y \mid X, Z\}$, simplificando novamente $P(y \mid x, t, z)$ por $P(y \mid x, z)$, assim obtemos:

$$\begin{aligned}
P(y \mid do(x)) &= \sum_t P(t) \sum_z P(z \mid x, t)P(y \mid x, t, z), \\
&= \sum_t P(t) \sum_z P(z \mid t)P(y \mid x, z), \\
&= \sum_z \sum_t P(t)P(z \mid t)P(y \mid x, z), \\
&= \sum_z \sum_t P(z, t)P(y \mid x, z), \\
&= \sum_z P(z)P(y \mid x, z).
\end{aligned}$$

□

Apesar da fórmula da porta dos fundos estender nossa capacidade de identificar efeitos causais, ela ainda possui limitações. Em especial, precisamos que exista pelo menos um conjunto de nós que cumpra as condições de porta dos fundos. Na próxima seção damos um exemplo onde não temos variáveis que façam papel de porta dos fundos e apresentamos uma nova ferramenta para identificação.

3.2.3 Porta da frente

Suponha que nos é dado o seguinte modelo causal sobre o efeito do fumo no pulmão com variáveis: se o paciente fuma, se desenvolveu câncer de pulmão e o se possui um certo gene ativo. O gene pode ou não ser responsável tanto pela vontade de fumar do indivíduo quanto pela predisposição de desenvolver câncer de pulmão, sendo assim uma variável que confunde o efeito do fumo do desenvolvimento de cancer.

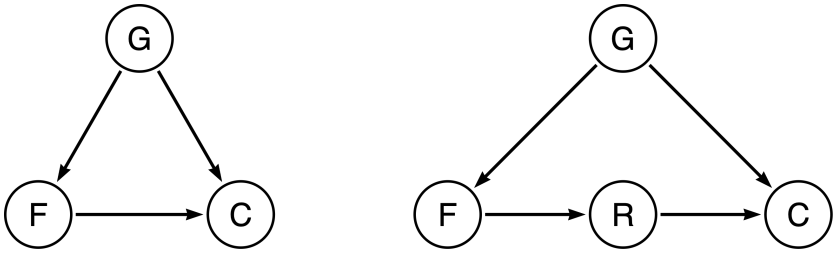

Figura 14: **Podemos modelar o efeito do fumo de duas formas, um modelo onde o tabaco tem um efeito direto no surgimento de câncer de pulmão (esquerda) e outro modelo onde o supomos que o efeito do fumo no pulmão é mediano pelo acúmulo de resíduos no pulmão**. Modelando da segunda forma podemos encontrar uma fórmula que identifica o efeito do tabaco na probabilidade de desenvolver câncer por meio de uma porta da frente.

Quando não temos acesso aos dados dos genes de cada indivíduo, o problema de identificar o efeito do cigarro no câncer é impossível de ser calculado. Se, no entanto, mudarmos nossa modelagem adicionando um novo elemento que pode ser medido, o resíduo acumulado no pulmão, e supormos que fumar só influencia no desenvolvimento de câncer a partir deste acúmulo, o efeito causal agora pode ser identificado mesmo sem a existência de uma porta dos fundos.

Para demonstrar isso identificaremos o efeito $P(C \mid do(F))$ no modelo 3.3. Sabemos pelo teorema da fatorização truncada 3.1 que podemos escrever a distribuição conjunta pós-intervenção como o produtório:

$$P(c, r, g \mid do(f)) = P(g)P(r \mid f)P(c \mid r, g).$$

Marginalizando na variável se o paciente desenvolveu câncer ou não, te-

mos:

$$
\begin{aligned}
P(c \mid do(f)) &= \sum_{r,g} P(g)P(r \mid f)P(c \mid r, g), \\
&= \sum_{r} P(r \mid f) \sum_{g} P(g)P(c \mid r, g).
\end{aligned}
$$

Observe que devido às independências geradas pela estrutura do modelo temos que:

$$\{G \perp\!\!\!\perp R \mid F\} \implies P(g \mid r, f) = P(g \mid f).$$

$$\{C \perp\!\!\!\perp F \mid G, R\} \implies P(c \mid r, f, g) = P(c \mid g, r).$$

Com isso simplificamos a expressão $P(c \mid do(f))$:

$$
\begin{aligned}
P(c \mid do(f)) &= \sum_{r} P(r \mid f) \sum_{f} \sum_{g} P(f)P(g \mid f)P(c \mid r, g), \\
&= \sum_{r} P(r \mid f) \sum_{f} \sum_{g} P(f)P(g \mid r, f)P(c \mid r, f, g), \\
&= \sum_{r} P(r \mid f) \sum_{f} \sum_{g} P(f)P(c, g \mid r, f), \\
&= \sum_{r} P(r \mid f) \sum_{f} P(f)P(c \mid r, f).
\end{aligned}
$$

Deixando o efeito causal em função de probabilidades condicionais conhecidas. Pearl, mostra que podemos interpretar esse efeito como a aplicação sequencias da fórmula da porta dos fundos [10, p. 82]. No primeiro passo identificamos o efeito $P(r \mid do(f)) = P(r \mid f)$. E em seguida identificamos o efeito $P(c \mid do(r)) = \sum_f P(f)P(c \mid f, r)$, utilizando a variável F como porta dos fundos. Dessa forma que temos que:

$$
\begin{aligned}
P(c \mid do(f)) &= \sum_r P(c, r \mid do(f)), && (1)\\
&= \sum_r P(r \mid do(f)) \cdot P(c \mid do(f), r), && (2)\\
&= \sum_r P(r \mid do(f)) \cdot P(c \mid do(f), do(r)), && (3)\\
&= \sum_r P(r \mid do(f)) \cdot P(c \mid do(r)), && (4)\\
&= \sum_r P(r \mid f) \sum_f P(f) \cdot P(c \mid r, f). && (5)
\end{aligned}
$$

Nos passos de 2 para 3 e de 3 para 4 fizemos transformações com o operador de *do* sem explicação, iremos justificá-los na próxima seção.

O método que utilizamos para resolver o problema de identificação do efeito do cigarro na probabilidade de desenvolver câncer de pulmão é uma instância de um teorema conhecido como fórmula da porta da frente. A ideia do ajuste pela porta da frente é encontrar um conjunto de variáveis Z que medeia todo o efeito da variável de tratamento na variável de interesse e que, simultaneamente, consiga ter o seu próprio efeito causal identificado por meio do ajuste pela porta dos fundos. Podemos formalizar esses pré-requisitos com o seguinte critério:

Definição 3.4 (**Critério da porta da frente**). *Dado uma DAG G e um par ordenado (X, Y) de nós pertencentes à G, um subconjunto $Z \subset G$ é chamado de porta da frente de (X, Y) se respeitar todas as seguintes 3 condições:*

1. *Z intercepta todos os caminhos direcionados entre X e Y.*
2. *Não existem caminhos dos fundos desbloqueados entre X e Z.*
3. *X bloqueia todos os caminho dos fundos entre Z e Y.*

Com o critério da porta da frente, podemos introduzir o seguinte teorema [10, p. 81]:

Teorema 3.4 (**Fórmula da Porta da Frente**). *Dado uma DAG $\mathcal{G}$ e um par ordenado (X, Y) de nós pertencentes à $\mathcal{G}$, se um subconjunto Z de variáveis de $\mathcal{G}$ for uma porta da frente de (X, Y) temos que:*

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{\hat{x}} P(\hat{x}) P(y \mid \hat{x}, z).$$

E novamente, apesar da fórmula da porta da frente estender nossa capacidade de identificar efeitos causais, somente com as três formulas, do ajuste, da porta dos fundos e da porta da frente, ainda não conseguimos identificar todos os efeitos causais quando eles são possíveis de serem identificados. Para isso precisamos de uma ferramenta ainda mais forte, conhecido como calculo de Do [10, p. 85].

3.3 Calculo de Do

Nessa seção estendemos nossa capacidade de identificar efeitos causais com um conjunto de três regras simples que nos permite fazer transformações simbólicas envolvendo o operador de *do*. Descobertas em [9], essas regras ficaram conhecidas como calculo de Do. Antes de apresentar as regras definimos as seguintes notações de grafos manipulados:

Definição 3.5 (**Grafos Manipulados**). *Dado uma DAG* $\mathcal{G} = (V, A)$ *e* X, Z, W *subconjuntos disjuntos de* V*, definimos:*

$\mathcal{G}_{\overline{X}}$ *o grafo* $\mathcal{G}$ *onde as arestas entre os nós de* X *e seus respectivos pais são deletadas.*

$\mathcal{G}_{\underline{X}}$ *o grafo* $\mathcal{G}$ *onde as arestas entre nós de* X *e seus respectivos filhos são deletadas.*

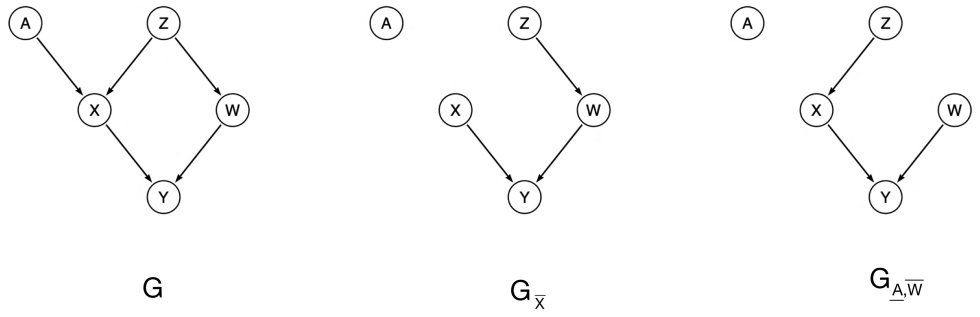

Figura 15: Na Imagem temos duas possíveis formas de manipular grafos, na esquerda temos o grafo original, sem manipulações, no meio temos o grafo manipulado que corresponde a intervir em X, e a direita manipulamos o grafo original removendo os mecanismos causais entre A e seu filho X e W e seu pai Z.

Teorema 3.5 (**Regras do cálculo de Do**). *Seja $\mathcal{B} = (\mathcal{G}, P)$ uma rede bayesiana causal, para quaisquer subconjuntos disjuntos $X, Y, Z, W \subset \mathcal{G}$ temos que:*

Regra 1 $P(y \mid do(x), z, w) = P(y \mid do(x), w)$ *se* $\{Y \perp\!\!\!\perp Z \mid X, W\}_{\mathcal{G}_{\overline{X}}}$.

Regra 2 $P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$ *se* $\{Y \perp\!\!\!\perp Z \mid X, W\}_{\mathcal{G}_{\overline{X}\underline{Z}}}$.

Regra 3 $P(y \mid do(x), do(z), w) = P(y \mid do(x), w)$ *se* $\{Y \perp\!\!\!\perp Z \mid X, W\}_{\mathcal{G}_{\overline{XZ(W)}}}$.

Onde $Z(W)$ são os nós pertencentes à Z que não possuem descendentes em W.

As regras do cálculo de Do foram provas completas [11] no sentido de que se o efeito causal é identificável então existe uma sequência finita de transformações obedecendo às regras do cálculo de Do que retorna uma expressão livre do operador de *do*. Uma consequência disso é que todas as regras derivadas até o momento podem ser recuperadas a partir do cálculo de Do.

3.3.1 Fórmula do ajuste revisitada

No começo do capítulo derivamos a fórmula do ajuste a partir do teorema da fatorização truncada, mostramos agora como podemos recuperar a mesma fórmula a partir das regras 1-3.

Prova 3.3 (**Fórmula do Ajuste**). *Tome $\mathcal{B} = (\mathcal{G}, P)$ uma rede bayesiana causal que fatoriza de acordo com $\mathcal{G} = (V, A)$. Queremos identificar o efeito $P(y \mid do(x))$ para $X, Y \in V$, tome $Z = Pa(X)$.*

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y, z \mid do(x)), \\ &= \sum_z P(y \mid do(x), z) \cdot P(z \mid do(x)). \end{aligned}$$

Como $\mathcal{B}$ é uma rede bayesiana, sabemos que $\{X \perp\!\!\!\perp (Desc(X))^c \mid Pa(X)\}_{\mathcal{G}}$, em particular, temos que $\{X \perp\!\!\!\perp Y \mid Z\}_{\mathcal{G}_{\underline{X}}}$ o que nos permite, pela regra 2, escrever $P(y \mid do(x), z) = P(y \mid x, z)$. Além disso, como $\mathcal{G}$ é acíclico, $Z = Pa(X) \not\subset Desc(X)$, logo temos que $\{X \perp\!\!\!\perp Z\}_{\mathcal{G}_{\overline{X}}}$, o que pela regra 3 nos garante que $P(z \mid do(x)) = P(z)$. Sendo assim temos que:

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y \mid do(x), z) \cdot P(z \mid do(x)), \\ &= \sum_z P(y \mid x, z) \cdot P(z). \end{aligned}$$

□

3.3.2 Identificabilidade

Sabemos que se não existir uma sequência finita de transformações que respeitem as regras do cálculo de Do para uma expressão da forma:

$$P(y \mid do(X = x), z).$$

Então o efeito não é identificável, no entanto, é importante que existam condições gráficas que nos garanta identificabilidade. Galles e Pearl [4], encontram quatro condições para que garantem que um efeito pode ser identificado quando X for um conjunto com um único elemento. Em particular,

eles demostram para uma classe mais geral de modelos dos que estudados até agora, conhecidos como modelos semi-markovianos.

Definição 3.6 (**Modelo Semi-markoviano**). *Um Modelo Semi-markoviano é uma rede bayesiana causal* $\mathcal{B} = (\mathcal{G}, P)$ *com variáveis observáveis* $V = \{v_1, v_2, ..., v_n\}$ *e variáveis latentes* $U = \{u_1, u_2, ..., u_k\}$ *onde cada uma das variáveis latentes são nós raízes e tenham no máximo dois filhos pertencentes a* V *de tal forma que* $U \cap V = \emptyset$ *e* $U \cup V = \mathcal{V}$ *onde* $\mathcal{V}$ *são os nós pertencentes à* $\mathcal{G}$.

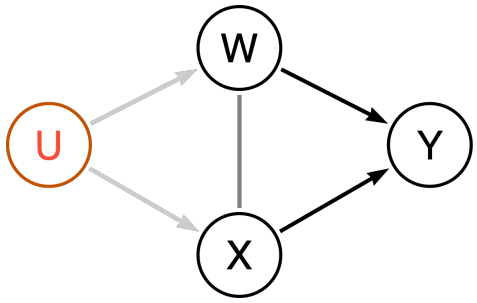

Figura 16: No modelo semi-markoviano estendemos as definições de uma RBC (3.2) para permitir variáveis latentes, representada pela cor vermelha, que confundem as variáveis observadas, criando correlações espúrias e dificultando a identificação de efeitos causais.

Teorema 3.6 (**Condições Suficientes: Galles e Pearl 1995**). *Seja um Modelo Semi-markoviano caracterizado por um grafo $\mathcal{G} = (V, A)$ e $X, Y \in V$, Satisfazer quais quer umas das quatro condições abaixo é condição suficiente para identificar o efeito $P(y \mid do(X = x))$:*

1. $\{X \perp\!\!\!\perp Y\}_{\mathcal{G}_{\overline{X}}}$
2. $Y \notin Desc(X)$
3. *Existe um conjunto de nós $Z \subset V$ tal que Z bloqueia todas as portas dos fundos entre X e Y de forma que $P(y \mid do(x))$ é identificável.*
4. *Existe conjuntos $Z_1, Z_2 \subset V$ tal que:*
   (a) $\{X \perp\!\!\!\perp Y \mid Z_1\}_{\mathcal{G}_{\overline{Z_1 X}}}$
   (b) $\{Z_1 \perp\!\!\!\perp Y \mid Z_2\}_{\mathcal{G}_{\underline{Z_1}\overline{X}}}$
   (c) $\{Z_1 \perp\!\!\!\perp X \mid Z_2\}_{\mathcal{G}_{\underline{X}}}$
   (d) $\{Y \perp\!\!\!\perp X \mid Z_1, Z_2\}_{\mathcal{G}_{\overline{Z_1 X(Z_2)}}}$

Além disso, Galles e Pearl [4] também apresentam a seguinte condição necessária para identificabilidade:

Teorema 3.7 (**Condição Necessária: Galles e Pearl 1995**). *Seja um Modelo Semi-markoviano caracterizado por um grafo $\mathcal{G} = (V, A)$ e $X, Y \in V$, caso nenhuma das condições do teorema 3.6 sejam satisfeitas para $X e Y$ então o efeito $P(y/middo(x))$ não é identificável.*

3.4 Intervenções Condicionais

Até o momento, estudamos intervenções da forma $do(X = x)$, entretanto é natural estender o estudo de efeitos causais não somente à simples intervenções determinísticas e atômicas e sim à possíveis políticas de intervenções, onde a ação feita na variável de tratamento dependa de outras variáveis que não sejam suas descendentes. Um exemplo disso é estudar o efeito de receitar um medicamento (D) ao paciente quando sua temperatura (T) atingisse certa temperatura na recuperação do paciente (R), $P(\mathrm{R} \mid do(\mathrm{D} = f(\mathrm{T})))$.

Mas como veremos, identificar o efeito de $P(y \mid do(X = x))$ não nos garante que consigamos identificar $P(y \mid do(X = g(z)))$, onde $g : Z \to X$ é

uma função qualquer. Provar isso é simples, tome uma RBC que fatora de acordo com $\mathcal{G} = (V, A)$ e $Y, X \in V$ e $Z \subset V$ tal que $Z \cap Desc(X) = \emptyset$. Para identificar o efeito de $P(y \mid do(X = g(Z)))$ expandimos:

$$P(y \mid do(X = g(Z))) = \sum_z P(y \mid do(X = g(z), z) \cdot P(z \mid do(X = g(z)).$$

Como $Z \cap Desc(X) = \emptyset$, não existe caminho direcionado entre X e Z, o que significa que $\{X \perp\!\!\!\perp Z\}_{\mathcal{G}_{\overline{X}}}$. Logo pela regra 3 temos que $P(z \mid do(X = g(Z))) = P(z)$, sendo assim simplificamos a expressão:

$$\begin{aligned} P(y \mid do(X = g(Z))) &= \sum_z P(y \mid do(X = g(z), z) \cdot P(z). \\ &= E_Z[P(y \mid do(X = g(z), z)] \end{aligned}$$

Correa e Bareinboim [3] generalizaram o cálculo de do para intervenções mais genéricas, possibilitando o estudo do efeito de planos mais condizentes com os que seriam aplicados na prática.

4 Contrafactuais

Nesse capitulo atingimos o pico da hierarquia causal, o estudo dos contrafactuais [10, ch. 7], do que não aconteceu, mas poderia ter acontecido. Em cada um dos capítulos desenvolvemos ferramentas capazes de responder perguntas cada vez mais difíceis. Começamos com questionamentos associativos da forma "Qual valor de Y esperamos encontrar dado que observamos X", as quais foram respondidas com modelos gráficos chamados de Redes Bayesianas, que nos possibilitava modelar nossos problemas introduzindo hipóteses de independências condicionais entre variáveis por meio de arcos em um grafo acíclico direcionado. Para que pudéssemos subir um nível na hierarquia de Pearl e responder perguntas causais da forma "O que podemos esperar de Y dado que X foi obrigado a tomar o valor x" com dados observacionais, tivemos que redefinir o significado da direção dos arcos nas Redes Bayesianas, dando origem às Redes Bayesianas Causais.

As perguntas de interesse desse capitulo não são mais simples intervenções, mas sim sobre intervenções no passado dado que sabemos o presente. Podemos imaginar que seria como se tivéssemos viajado em uma máquina do tempo e interferido no passado. As perguntas contrafactuais são da forma "O que poderíamos ter esperado de Y se tivéssemos obrigado X a tomar o valor x, mas observamos y e $\hat{x}$". No primeiro momento, o estudo de contrafactuais pode ser, erradamente, visto como algo sem valor prático, qual a utilizada de estudar aquilo que não aconteceu, mas poderia ter acontecido?

4.1 Onde Encontramos Contrafactuais?

Apesar de contrafactuais parecerem complicados demais e muito distantes da realidade, não é difícil encontrar aplicações relevantes. Um ambiente onde contrafactuais aparecem naturalmente é no meio jurídico. Perguntas como "X é o culpado por Y" podem ser reformuladas como "Se X tivesse agido diferente, teria Y tido um resultado diferente do que observamos?".

Outro cenário onde contrafactuais são uteis é no processo de *troubleshooting*. Na produção de código, por exemplo, é comum esbarramos em *bugs*, onde nosso código retorna algo inesperado. Nesses momentos o programador se pergunta onde foi que ele errou para ter chegado até aqui e o que ele pode fazer (ou deveria ter feito) para corrigir seu programa. Uma das primeiras aplicações do cálculo de contrafactuais foi no sistema, da Microsoft, Windows 95 [6], onde eles usavam modelos contrafactuais para encontrar e

consertar problemas como falha na impressão ou na demora para abrir algum programa.

4.2 Modelos Estruturais Causais

Os Modelos Estruturais Causais, mais conhecidos como *Structural Causal Models* (SCM) são as estruturas-base para o estudo de contrafactuais. Diferente das RBC que foram uma extensão natural das Redes Bayesianas, definiremos as SCM do zero e em seguida mostraremos como RB e RBC podem ser recuperadas naturalmente dentro das SCM.

Definição 4.1 (**SCM markoviana**). *Um SCM* $\mathcal{M}$ *é uma tupla* $(\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$ *onde:*

$\mathbf{U} = \{U_1, ...U_n\}$ *é o conjunto de todas as variáveis exógenas*[8].

$\mathbf{V} = \{V_1, ...V_n\}$ *é o conjunto de todas as variáveis endógenas*[9].

$\mathcal{F} = \{f_1, ...f_n\}$ *é o conjunto de funções tais que* $f_i : U_i \cup \mathbf{V} \setminus V_i \rightarrow V_i$*, de tal forma que* $\mathcal{F}$ *mapeia* $\mathbf{U}$ *em* $\mathbf{V}$*.*

$P(\mathbf{U})$ *é uma função de probabilidade definida sob o domínio de* $\mathbf{U}$ *tal que* $\{U_i \perp\!\!\!\perp U_j\}, \forall i \neq j$*.*

Ao enfraquecer a suposição de independência entre as variáveis exógenas definimos uma classe mais geral chamada de SCM Semi-markovianas.

Em muitos contextos as variáveis exógenas representam unidades, informação no nível dos indivíduos presentes no modelo. Para exemplificar pegue o mesmo cenário visto no capítulo passado 3.1, onde estávamos estudando o efeito de dois medicamentos, as variáveis endógenas do nosso modelo são: droga administrada, idade do paciente e resultado do tratamento. A variável exógena U_{idade} pode representar o próprio individuo sendo tratado no momento, ou todo conjunto de informação que o identifica unicamente.

[8]Variáveis exógenas são aquelas as quais seus comportamentos são determinados por variáveis que não contempladas no modelo

[9]variáveis endógenas são aquelas as quais seus comportamentos são totalmente determinados por outras variáveis presentes no modelo

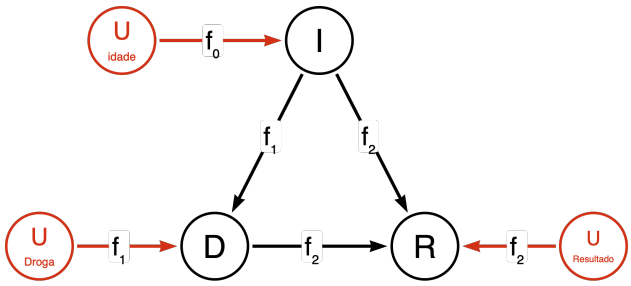

Figura 17: **Extensão do modelo que estuda o efeito das drogas A e B na recuperação dos pacientes (3.2.1)**. Aqui adicionamos as variáveis latentes e especificamos o comportamento funcional para permitir contrafactuais. As variáveis latentes representam as unidades, ou indivíduos do modelo.

4.2.1 SCM no Primeiro Degrau Causal: Observações

Como os valores de $\mathbf{V}$ são unicamente determinados a partir de $\mathbf{U}$, $P(\mathbf{U})$ induz naturalmente uma distribuição sob $\mathbf{V}$. Dada uma SCM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$ e $\mathbf{Y} \subset \mathbf{V}$ definimos a distribuição induzida em $\mathbf{V}$ como:

$$P^{\mathcal{M}}(\mathbf{y}) := \sum_{\{\mathbf{u}|\mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u})$$

Onde $\mathbf{Y}(\mathbf{u})$ é o valor das variáveis em $\mathbf{Y}$ após avaliar $\mathcal{F}(\mathbf{u})$. Veja que com a distribuição induzida recuperamos a distribuição conjunta das variáveis endógenas do modelo, então perguntas do primeiro nível causal podem ser respondidas utilizando SCM.

Pegue o exemplo da figura acima, no capítulo passado, ao utilizar RBC, especificaríamos a distribuição conjunta das variáveis do modelo como algo da forma:

$$I \sim P_1.$$
$$D \mid I \sim P_2(I).$$
$$R \mid I, D \sim P_3(I, D).$$

Onde P_1, P_2, P_3 são distribuições quaisquer. O cálculo de qualquer pergunta do primeiro nível poderia ser facilmente respondido com a distribuição conjunta que é recuperada facilmente a partir de sua fatorização.

$$P(I, D, R) = P_1(I)P_2(D \mid I)P_3(R \mid D, I).$$

Nas SCM, a incerteza não é modelada nas variáveis endógenas, elas são determinísticas dado as variáveis exógenas. Logo, as distribuições são definidas somente para variáveis exógenas, e a partir das relações funcionais, recuperaríamos a distribuição conjunta das variáveis observáveis.

$$\begin{aligned} U_I &\sim P_1', \\ U_D &\sim P_2', \\ U_R &\sim P_3', \\ I &\sim f_0(U_I), \\ D &\sim f_1(U_D, I), \\ R &\sim f_1(U_R, I, D). \end{aligned}$$

A distribuição conjunta das variáveis endógenas fica menos explicita, é o preço que se paga por incluir hipóteses sobre o comportamento funcional das variáveis.

4.2.2 SCM no Segundo Degrau Causal: Intervenções

Fica claro então que com SCM conseguimos formular perguntas do primeiro nível da hierarquia de Pearl. Além disso, conseguimos formular perguntas intervencionistas com SCM utilizando submodelos, SCM manipulas, onde redefinimos as funções em $\mathcal{F}$ para simular uma intervenção.

Definição 4.2 (**Sub-Modelo de uma SCM**).
Dada uma SCM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$, um subconjunto $\mathbf{X} \subset \mathbf{V}$ e uma realização $\mathbf{x}$ de $\mathbf{X}$, $\mathcal{M}_{\mathbf{x}}$ é um submodelo de $\mathcal{M}$ definido como:

$$\mathcal{M}_{\mathbf{x}} = (\mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U})).$$

Onde $\mathcal{F}_{\mathbf{x}} = \{f_i \mid V_i \notin X\} \cup \{f : \mathbf{X} \to \{\mathbf{x}\}\}$

Com a noção de submodelos podemos definir probabilidades de eventos dado uma intervenção em um subconjunto das variáveis endógenas. Dado uma SCM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$ e $\mathbf{Y}, \mathbf{X} \subset \mathbf{V}$, definimos $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ como sendo o valor das variáveis $\mathbf{Y}$ após avaliar $\mathcal{F}_{\mathbf{x}}$ em $\mathbf{u}$, ou seja $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$. Desta forma a SCM induzem naturalmente uma distribuição conjunta para cada intervenção $\mathbf{x}$ definida como:

$$P^{\mathcal{M}}(\mathbf{Y_x} = \mathbf{y}) = \sum_{\{\mathbf{u}|\mathbf{Y_x}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u})$$

Para exemplificar vamos criar um sub-modelo da SCM do exemplo do tratamento das drogas.

Figura 18: **A esquerda temos a SCM original enquanto a direita temos seu sub-modelo onde substituímos a f_1 por x**. A substituição da função por um valor fixo pode ser vista como uma intervenção onde substituímos o mecanismo de geração natural do modelo por um valor artificial.

No sub-modelo que representa uma intervenção no tipo de tratamento, a nova distribuição das variáveis endógenas podem ser recuperadas como:

$$\begin{aligned}
&U_I \sim P_1^{'},\\
&U_D \sim P_2^{'},\\
&U_R \sim P_3^{'},\\
&I \sim f_0(U_I),\\
&D = x,\\
&R \sim f_1(U_R, I, x).
\end{aligned}$$

De tal forma que a probabilidade de um paciente se recuperar após forçar um tratamento x, $P^{\mathcal{M}_{\mathbf{x}}}(R = 1)$, pode ser calculada como:

$$P^{\mathcal{M}_{\mathbf{x}}}(R = 1) = \sum_{\{\mathbf{u}|\mathbf{f_1}(\mathbf{U_I},\mathbf{I},\mathbf{x})=\mathbf{1}\}} P(U_I, U_D, U_R) = \sum_{\{U_R,U_I\}} P(f_1(U_R, f_0(U_I)) = 1)$$

Toda SCM $\mathcal{M}$ pode ser associada à uma DAG $\mathbf{G}(\mathcal{M})$, onde as variáveis $U_i \cup \mathbf{V} \setminus V_i$ apontam para V_i. $\mathbf{G}(\mathcal{M})$ é chamado de diagrama causal de $\mathcal{M}$, e

assim como as RBC, suas arrestas representam um mecanismo casual. A similaridade entre SCM e RBC é evidente [2], em particular RBC exitem dentro SCM. O Teorema 4.1 possibilita que utilizemos todos os resultados derivados no capitulo três. Sendo assim, podemos utilizar a fórmula do ajuste, porta dos fundos, porta da frente e até mesmo o cálculo de Do ao trabalharmos com SCM.

Teorema 4.1 (SCM-RBC)**.** *Dado uma SCM* $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$*, seu diagrama causal* $\mathbf{G}(\mathcal{M})$ *é uma Rede Bayesiana Causal de acordo com* $P_{\mathcal{M}}$*.*

Como essa conexão entre SCM e RBC podemos escrever $P(\mathbf{Y_x} = \mathbf{y})$ como $P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x}))$ [2]. Agora que sabemos calcular perguntas dos dois primeiros níveis da hierarquia causal com SCM, nos resta calcular contrafactuais, observaremos na próxima seção que tudo que precisamos já foi definido.

4.2.3 SCM no Último Degrau Causal: Imaginando Contrafactuais

O cálculo de contrafactuais é inspirado na forma como imaginamos. Como não existe experimento que consiga responder, como seria o presente se o passado tivesse sido outro, precisamos imaginar. O processo de responder perguntas da forma: "Qual o valor esperado de Y se tivéssemos feito x, mas na realidade observamos y e $\hat{x}$", pode ser decomposto em três partes:

1. **Abdução**: Atualizamos nossas crenças sobre os fatores exógenos com base no que observamos no mundo factual.

2. **Ação**: Imaginamos como nossa intervenção muda os mecanismos causais no modelo de mundo que temos.

3. **Predição**: Utilizamos nossas novas crenças no novo modelo de mundo para imaginar o que poderia ter acontecido.

Uma vantagem das SCM é que elas nos permitem replicar esses passos de forma simples. Dado uma SCM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$ podemos definir a probabilidade de eventos contrafactuais $\mathbf{Y_x}, \ldots \mathbf{Z_w}$ para quaisquer subconjuntos $\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{W} \subset \mathbf{V}$ como:

$$P^{\mathcal{M}}(\mathbf{Y_x} = \mathbf{y}, \ldots \mathbf{Z_w} = \mathbf{z}) = \sum_{\{\mathbf{u} \mid \mathbf{Y_x}(\mathbf{u}) = \mathbf{y}, \ldots \mathbf{Z_w}(\mathbf{u}) = \mathbf{z}\}} P(\mathbf{u})$$

.

Com isso, expressões como $P(y_x \mid y_2, x_2)$ representam perguntas da forma "Qual a probabilidade de Y ter sido y se tivéssemos feito x mas o que realmente aconteceu foi y_2 e x_2". Os três passos que descrevemos podem ser formalizados no cálculo de contrafactuais com o seguinte teorema [10, cap. 7]:

Teorema 4.2 (Cálculo de Contrafactuais). *Dado uma SCM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$, a expressão $P(\mathbf{Y_x} \mid \mathbf{W})$ pode ser calculada seguindo:*

1. ***Abdução**: Atualizar $P(\mathbf{U})$ com $P(\mathbf{U} \mid \mathbf{W})$*
2. ***Ação**: Modicar $\mathcal{M}$ com a intervenção $\mathbf{x}$, obtendo $\mathcal{M}_{\mathbf{x}}$.*
3. ***Predição**: Utilizar a SCM ${\mathcal{M}^*}_{\mathbf{x}} = (\mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U} \mid \mathbf{W}))$ para calcular $P^{\mathcal{M}^*\mathbf{x}}(\mathbf{Y}.) = P(\mathbf{Y_x} \mid \mathbf{W})$*

Continuamos com o exemplo do estudo do efeito da droga A e B, mas agora imaginando um presente alternativo. Após um tratamento bem-sucedido com a droga A, um paciente idoso pergunta "Como eu estaria se tivesse tomado a droga B?". SCMs nos dão as ferramentas para podermos responder tais perguntas sistematicamente. Utilizando o teorema 4.2 dividimos o cálculo da probabilidade contrafactual em 3 etapas:

1. Dado o que observamos, que o paciente é idoso, que ele se recuperou e que a droga utilizada foi a B, podemos atualizar P(B) como:

$$
\begin{aligned}
P(\mathbf{U} \mid \mathbf{V}) &= P(U_I, U_D, U_R \mid I = 1, D = A, R = 1), \\
&= P(U_I, U_D, U_R \mid I = 1, D = A, R = 1), \\
&= \frac{P(U_I, U_D, U_R, I = 1, D = A, R = 1)}{P(I = 1, D = A, R = 1)}, \\
&= \frac{P(U_I, U_D, U_R)}{P(I = 1, D = A, R = 1)}, \\
&= \frac{P(U_I, U_D, U_R)}{P(I = 1, D = A, R = 1)}, \\
&= \frac{P(U_I, U_D, U_R)}{\sum_{\{\mathbf{U} \mid \mathcal{F}(\mathbf{U}) = (1, A, 1)\}} P(U_I, U_D, U_R)}.
\end{aligned}
$$

2. Seguimos Modificando a SCM trocando f_1 por B, o equivalente a forçar o tratamento com a droga B.

3. Após incluir a intervenção na SCM, atualizamos $P(\mathbf{U})$ por $P(\mathbf{U} \mid \mathbf{V})$ e repetimos o cálculo do exemplo anterior.

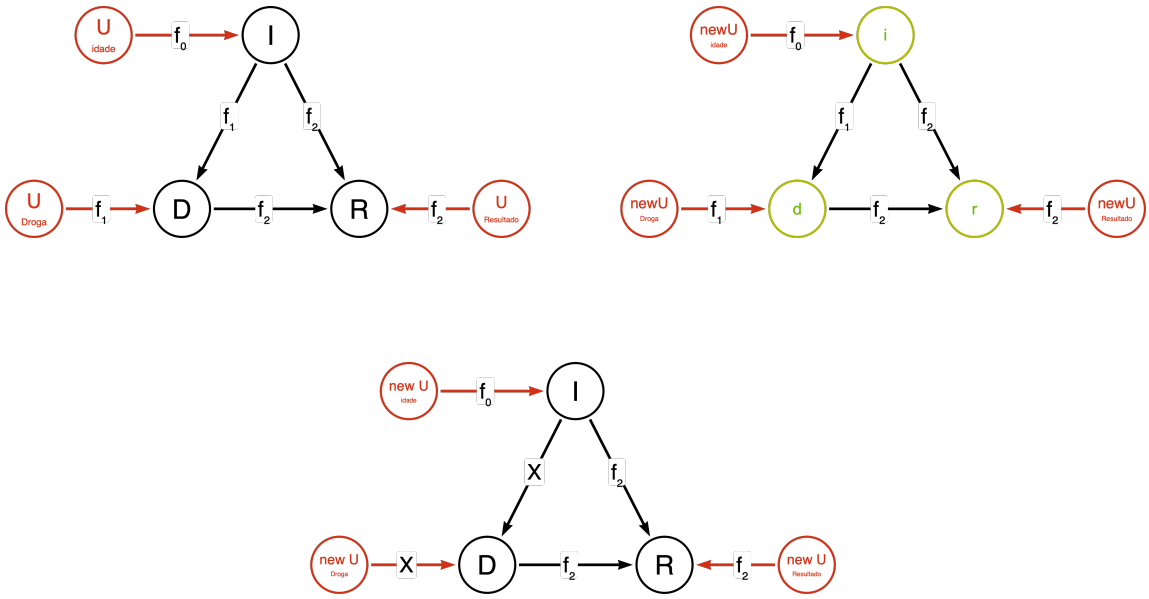

Figura 19: **As três etapas do cálculo de probabilidades contrafactuais.** Começamos com a SCM original do modelo (superior esquerda), após observar as variáveis endógenas podemos atualizar as probabilidades de $P(U)$ para $P(U \mid V)$. Por fim utilizamos a probabilidade atualizada de U no submodelo que representa o sistema com intervenção para calcular a probabilidade de contrafactuais.

Apesar de já conseguirmos calcular probabilidades de eventos contrafactuais, as etapas descritas podem se tornar confusas na prática. Balke e Pearl [1] propõem uma representação que facilita o processo. A ideia é representar graficamente as variáveis contrafactuais como sendo nós que compartilham as mesmas variáveis exógenas de suas respectivas variáveis factuais. Dessa forma obtemos um grafo espelhado chamado de grafo gêmeo como abaixo:

Com grafos gêmeos, perguntas como se $\{A_x \perp\!\!\!\perp C \mid B\}$ são facilmente respondidas. Repare que factualmente $\{A \perp\!\!\!\perp C \mid B\}$ é claramente verdade,

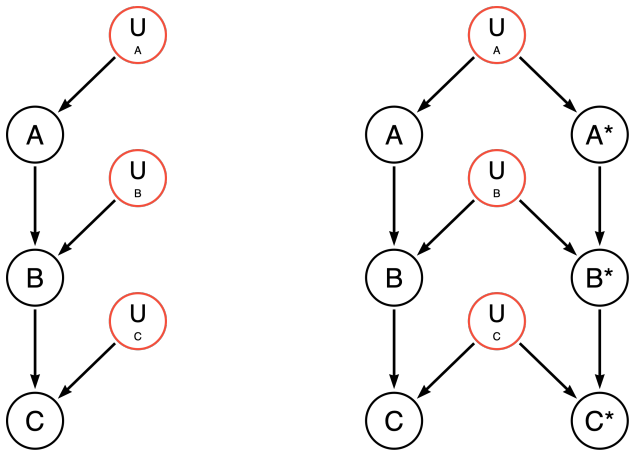

Figura 20: **No lado esquerdo temos um SCM e a esquerda temos uma representação do mesmo SCM no formato de um grafo gêmeo**. No grafo gêmeo é mais fácil de estudar graficamente quando efeitos causais são identificáveis entre o universo factual e contrafactual.

mas não é isso que estamos buscando saber, queremos identificar se A carrega informação sobre como C poderia ter sido caso tivéssemos feito x, mesmo já tendo informação sob B. Mesmo o caminho $A \to B \to C$ estando bloqueado por B, o caminho $A \to B \leftarrow U_B \to B^* \to C^*$ é liberado ao condicionarmos em B, um efeito comum. Dessa forma, informação consegue fluir entre universos contrafactuais e factuais.

Com isso encerramos os estudos de contrafactuais. A revisão de contrafactuais não é exaustiva, existem muitos assuntos que não foram comentados nesse capitulo, para um estudo mais completo recomendo a leitura de *Causality* [10] do pai da causalidade, Judea Pearl. Uma leitura mais leve pode ser realizada em *causal inference in statistics* [5]

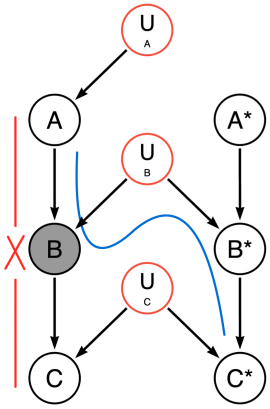

Figura 21: **O caminho factual $A \to B \to C$ é bloqueado por B, contudo o caminho contrafactual $A \to B^* \to C^*$ não está bloqueado**.

5 Conclusão

Neste trabalho escalamos cada degrau da hierarquia causal de Pearl. Começamos com o estudo das perguntas observacionais da forma: "Dado que observamos, X o que podemos esperar de Y?". Para responder esses tipos de pergunta bastaria ter conhecimento da distribuição conjunta das variáveis do modelo. Contudo, quanto maior do número de variáveis, mais difícil se torna a inferência dos parâmetros do modelo e mais dados são necessários [8]. Uma forma que encontramos de contornar esse problema é adicionando hipóteses de independências condicionais entre variáveis. Isso permitia fatorar a distribuição de probabilidade conjunta do modelo, o que reduzia o espaço de busca dos parâmetros a serem inferidos. Essa fatorização induzia naturalmente um grafo direcionado acíclico (2.7), o que nos motivou a definição de Redes Bayesianas (3.2).

O uso das Redes Bayesianas era adequado para perguntas observacionais, mas assim que começamos a fazer perguntas sobre tomadas de decisão esbarramos em um famoso problema conhecido como paradoxo de Simpson. A melhor decisão parecia depender de como olhávamos para os dados. A causa do paradoxo [12] estava escondida na famosa frase *"correlação não implica em causalidade"*. Visando subir um degrau na hierarquia causal, mudamos a interpretação dos arcos das Redes Bayesianas para representarem agora o mecanismo causal entre variáveis. A nova interpretação das arestas nos di-

recionou à definição de Redes Bayesianas Causais. Com a nova estrutura em mãos, perguntas da forma: "Se fizermos X o que podemos esperar de Y", poderia ser formulada a partir de uma manipulação do grafo original. Onde a intervenção em X poderia ser representada pela retirada dos mecanismos causais entre X e seus pais, obtendo assim um grafo mutilado. Peguntas causais do grafo original poderiam ser feitas como perguntas observacionais no novo grafo. O estudo principal no estudo das Redes Bayesianas Causais se tornou como escrever expressões probabilísticas formuladas no grafo manipulado em termos de probabilidades do grafo original, que é o que temos acesso na prática. Com esse objetivo de identificar efeitos causais a partir de dados observacionais, derivamos a fórmula do ajuste (3.2), a fórmula da porta dos fundos e a fórmula da porta da frente [10]. Nenhum dos três métodos, no entanto, garantem a identificação de efeitos causais mesmo se eles existirem. Para estender a capacidade de identificação apresentamos o Cálculo de Do [10]. Um conjunto de três regras simbólicas que permite transformar observações em intervenções e intervenções em observações quando certas hipóteses no grafo do modelo são cumpridas. Além do cálculo de Do conseguir recuperar todas as fórmulas, do ajuste, da porta dos fundos e da porta da frente, ela também é completa no sentido de que se existe o efeito é identificável, então existe uma sequência de transformações do cálculo de Do que identifica o efeito causal.

Por fim atingimos o terceiro e ultimo degrau da hierarquia de Pearl, o estudo dos contrafactuais. Perguntas contrafactuais questionam como seria o presente se o passado tivesse sido diferente. Para conseguir formular e fazer sentido de perguntas contrafactuais, abandonamos momentaneamente as Redes Bayesianas Causais e definimos os *Structural Causal Models* e mostramos que tanto as perguntas do primeiro nível quanto do segundo nível poderiam ser formuladas nesse novo modelo, e o motivo disso é que os SCM induzem Redes Bayesianas causais. Logo em seguida mostramos como fazer sentido de perguntas contrafactuais por meio de três etapas: Abdução, Ação e Predição. Esse processo se baseia em atualizar as crenças que temos sobre fatores exógenos com base com o que aconteceu no mundo factual, e utilizar essas novas crenças para prever como seria o mundo se tivéssemos interferido no passado. Como ferramenta que ajuda a estudar perguntas contrafactuais, introduzimos os grafos gêmeos, Redes Bayesianas que modelam, em simultâneo, o universo factual e os universos contrafactuais e os liga por meio de variáveis exógenas que seriam idênticas em ambos universos. Com isso completamos a travessia pelos degraus da causalidade, entendo como

prever (Redes Bayesianas), intervir (Redes Bayesianas Causais) e imaginar (SCM).

Referências

[1] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994.

[2] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. 2022.

[3] Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100, 2020.

[4] David Galles and Judea Pearl. Testing identifiability of causal effects. *arXiv preprint arXiv:1302.4948*, 2013.

[5] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[6] David Heckerman, John S Breese, and Koos Rommelse. Decision-theoretic troubleshooting. *Communications of the ACM*, 38(3):49–57, 1995.

[7] Guido W Imbens and Donald B Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

[8] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[9] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[10] Judea Pearl. *Causality*. Cambridge university press, 2009.

[11] Ilya Shpitser and Judea Pearl. *Identification of joint interventional distributions in recursive semi-Markovian causal models*. eScholarship, University of California, 2006.

[12] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[13] Jan Sprenger and Naftali Weinberger. Simpson's Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.