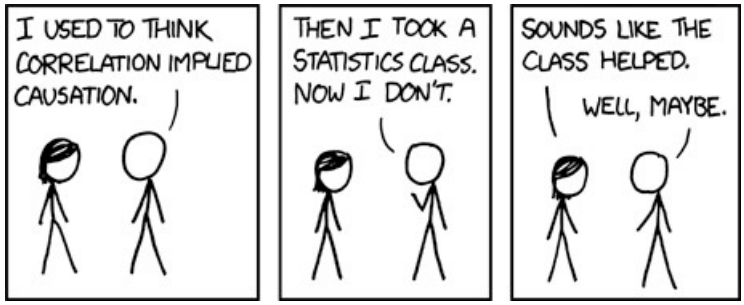
Causalidade - DAG

Manoel Galdino

CAUSALIDADE

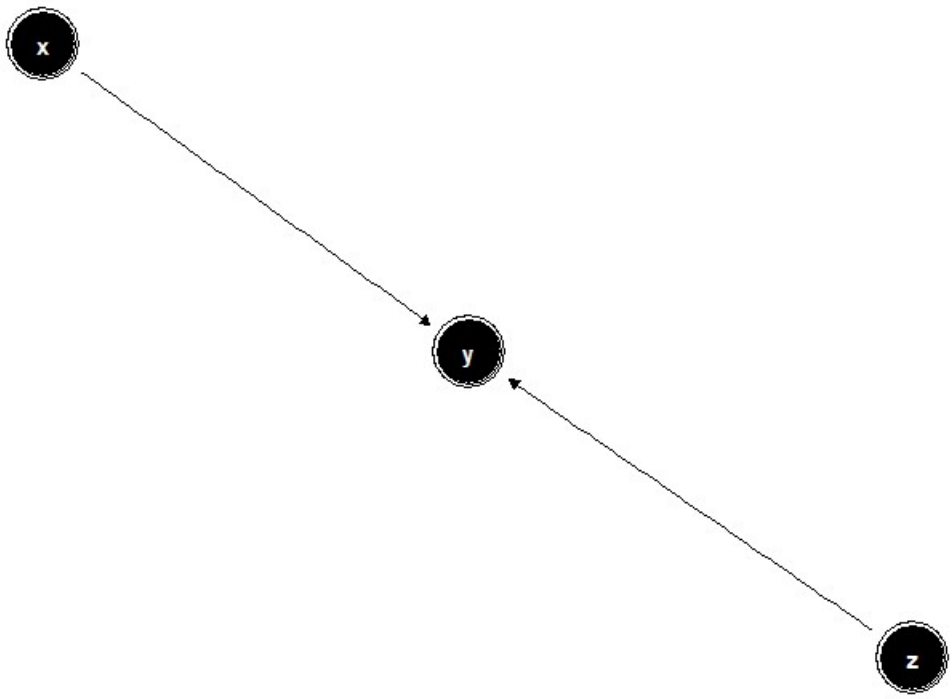


Fonte: xkcd

Introdução

Uma das principais abordagens para fazer inferência causal utiliza diagramas causais, chamados de Directed Acyclic Graphs (DAG). Ela foi desenvolvida na ciência da computação entre os anos 80 e 90 e é associada com o trabalho pioneiro de Judea Pearl. Veja o livro The Book of Why para um história de como surgiu essa abordagem.

Abaixo temos um exemplo simples de um DAG



Eles são chamados de DAG porque os gráficos são direcionados (apontam em uma direção), acyclic, porque não

permitem ciclos (isto é, se A causa B, B não pode causar A), e graphs porque, como você pode imaginar, são gráficos.

No exemplo acima, o DAG é formado por três variáveis {y, x, z}, que são em geral variáveis aleatórias. E as flechas indicam direção de causalidade. Ou seja, x causa y e z causa x. É importante saber que DAGs são não paramétricos. Eles podem ser interpretados como: y = f(x,z). Ou seja, qualquer função de x e z são igualmente possíveis. Eis alguns exemplos compatíveis com o DAG acima:

y = x + z

y = 10 + x + z + x*z

y = 3*x^z

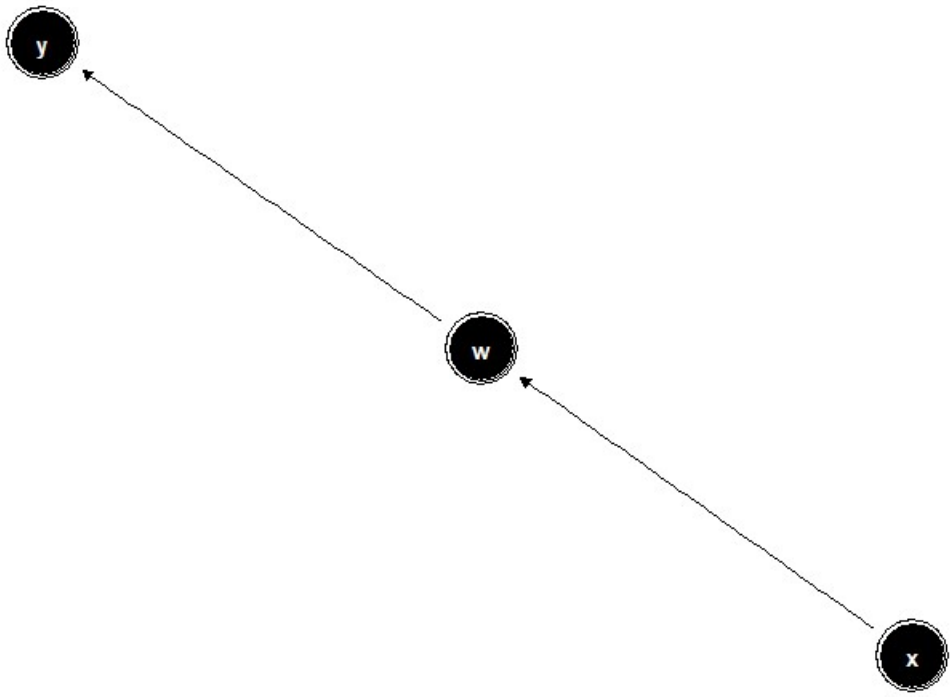
y = pi*z/x + x^2 + 1/(z^3)

A razão porque não escrevemos DAGs como equações é porque y = f(x,z) não expressa adequadamente a relaçao de causalidade, pois em matemática é a tanto faz escrever f(x,z) = y ou y = f(z,x). Porém, dizer que x e z causam y é muito diferente de dizer que y causa x e z. E com o DAG, as flechas indicam a direção da causalidade.

Os tipos básicos de DAGs

Temos basicamente três tipos de DAGs

1. chains

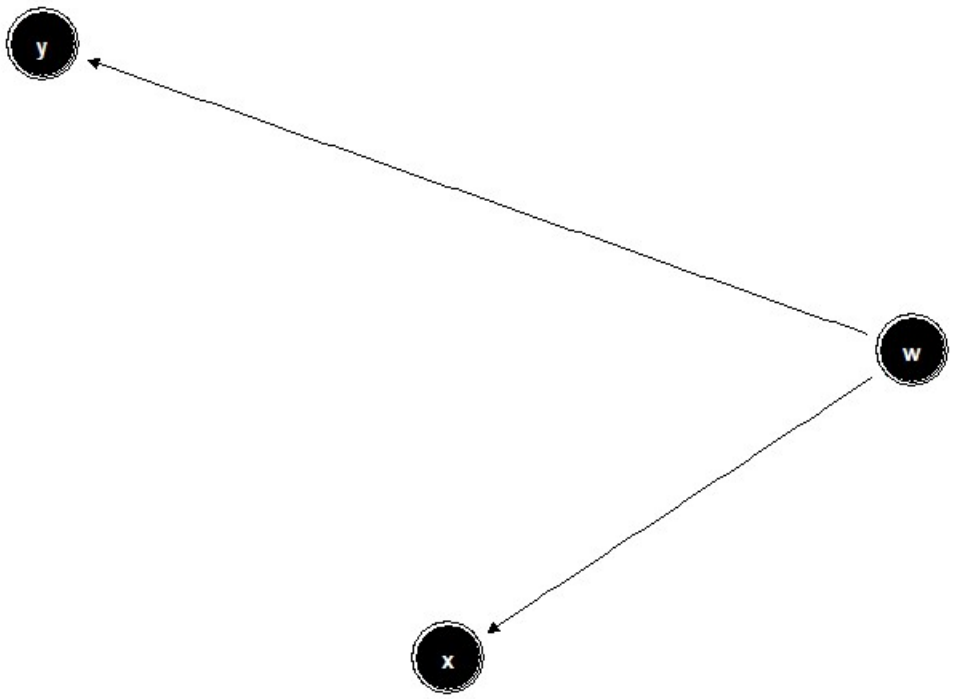


Aqui, x causa w que causa y. w pode ser pensado como mediador do efeito de x sobre y, isto é, x causa y via w.

Um exemplo simples seria que o desempenho econômico de um país aumenta a popularidade do presidente, que

causa mais votos. Ou seja, economia -> popularidade -> voto.

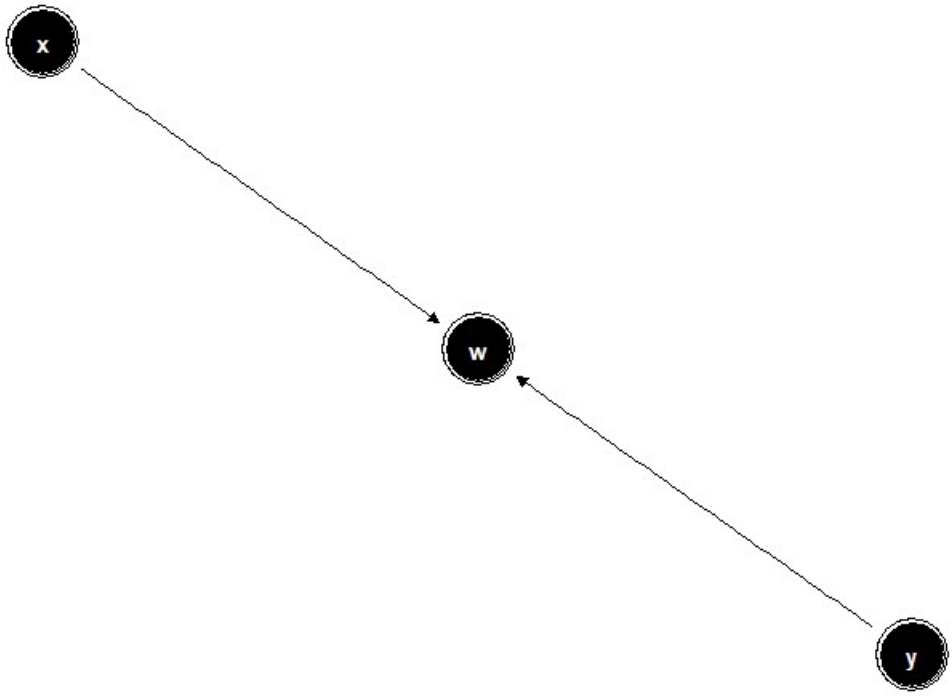
2. Forks



Aqui, w causa ambos x e y. w pode ser pensado como uma causa comum de x e y. Como veremos depois, esse tipo de gráfico é quase sempre o que as pessoas têm em mente quando falam de correlação espúria.

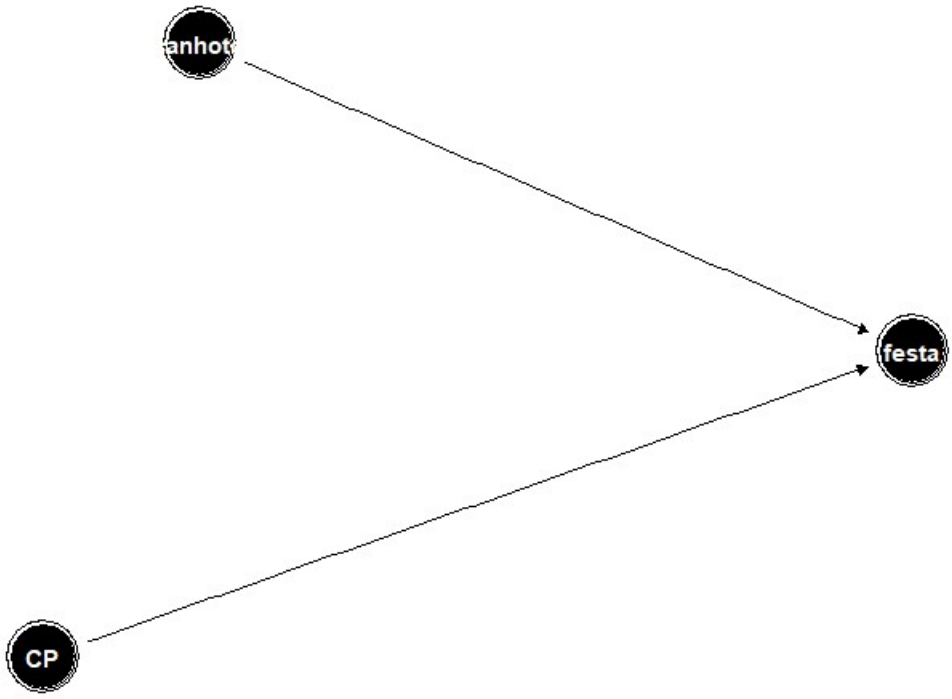
Um exemplo simples seria que "qualidade" de um candidato faz com que ele arrecade mais dinheiro para campanha eleitoral e também obtenha mais voto. Ou seja, qualidade -> constribuições de campanha e qualidade -> voto.

3. Colliders



Aqui, x causa w e y causa w. É também chamaood de fork (garfo) invertido. É claro pelo gráfico que não há relação causal entre x e y. Porém, como ambos causam w, uma análise que controle para w pode introduzir correlação onde na verdade não existe.

Para pensar em um exemplo de Collider, imagine que eu vou dar uma festa, e convido para ela apenas pessoas que fazem ciência política ou pessoas canhotas. Vamos supor que não há relação na populaçao em geral entre faer ciência política e ser canhota. Na minha festa, se a pessoa é canhota, ela deve fazer ciência política?



Vamos rodar no R uma simulação para ilustrar o collider bias. Vou supor que 10% das pessoas fazem ciência política e 5% são canhotas.

```
library(dplyr)
 set.seed(4)
 cp <- rbinom(1000, 1, p=.1)
 canhoto <- rbinom(1000, 1, p = .05)
 festa <- ifelse(cp == 1, 1,
                 ifelse(canhoto == 1, 1, 0))
 tabela <- data.frame(cp, canhoto, festa)
 round(cor(cp, canhoto),2)
```

```
## [1] -0.02
```

```
 tabela %>%
   filter(festa == 1) %>%
   summarise(round(cor(cp, canhoto),2))
```

```
##   round(cor(cp, canhoto), 2)
## 1                      -0.95
```

Na população em geral, a correlação é próxima de zero (-0,02), enquanto que entre as pessoas que foram à festa, a corelação é de -0,95! Condicionar entre quem foi na festa induz correlação espúria.

Definições

Path (caminho) é uma sequência de flechas vizinhas, Um directed path (caminho dirigido) é um path de sempre na mesma direção (x -> z -> y)

Relações entre variáveis (nós)

As relações entre variaveis são descritas por termos que usamos em genética: pais, filhos, ancestrais, descendentes e vizinhos. Os termos são intuitivos, mas pais e filhos referem-se a relações diretas (sem intermediários) entre variáveis, ancestrais e descendentes são variáveis em qualquer lugar no caminho (path) de ou para uma variável (nó).

Um path sem collider está aberto Um path com collider está fechado.

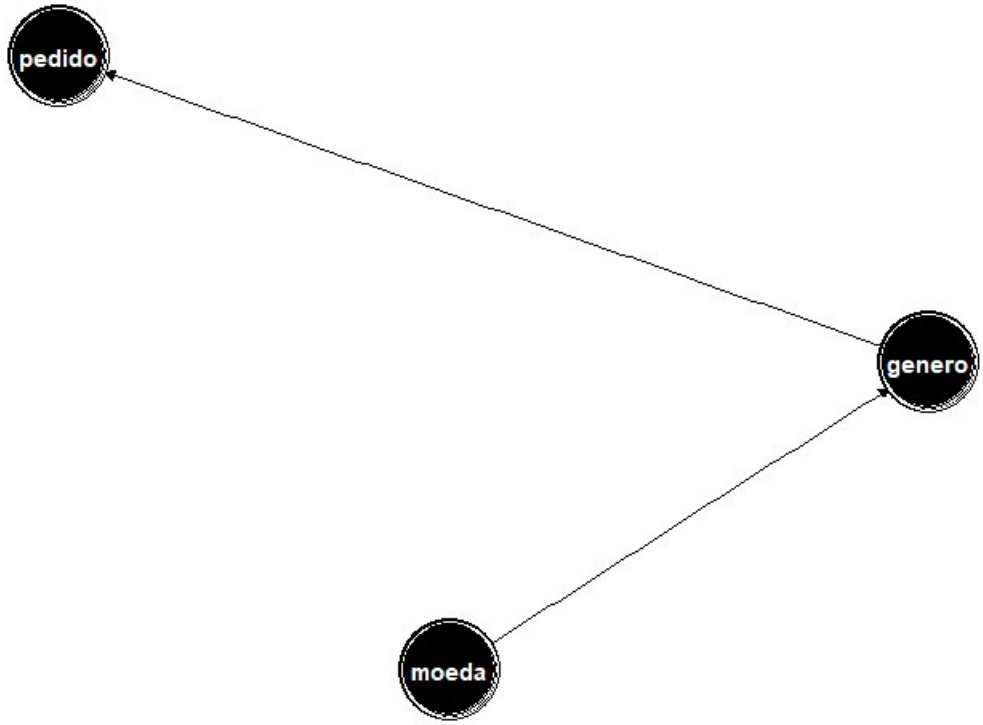
Duas variáveis (ou conjuntos de variáveis) estão d-separated se não há caminho aberto entre elas. Se há caminho aberto, pode ou não haver independência (vários caminhos abertos podem se cancelar)

Controle

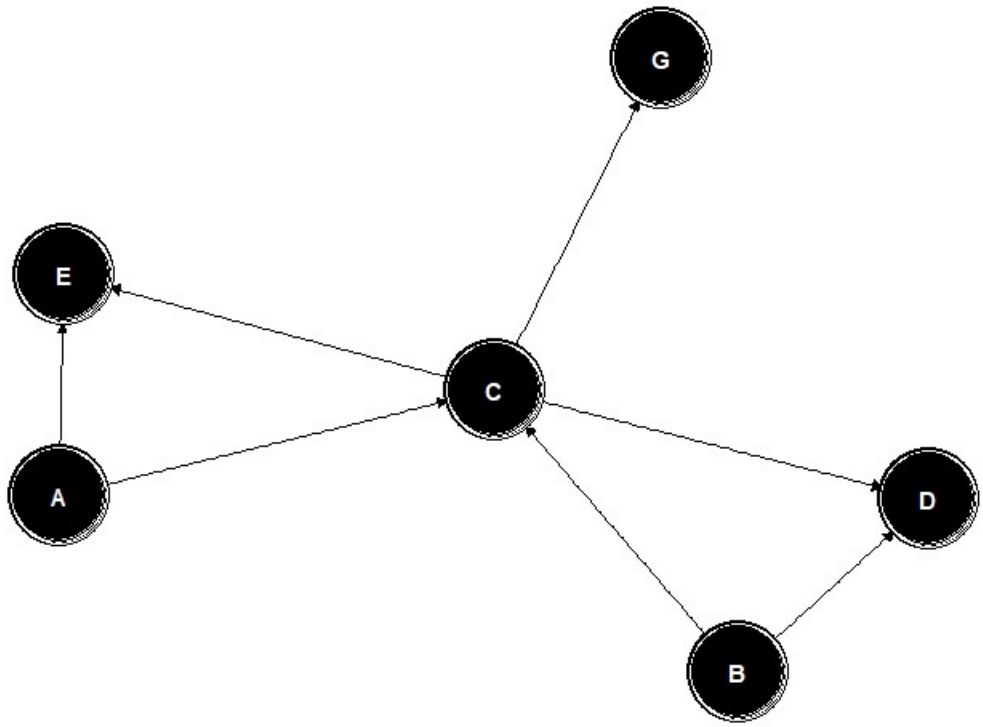
Falamos o tempo inteiro em controlar para variáveis. Em experimentos, controlar significa manipular o valor das

variáveis. Em estudos observacionais, controlar significa condicionar (estratificar ou colocar em uma regressão)

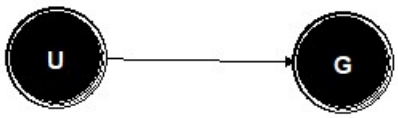
Em termos de DAG, isso sinifica: 1. Manipular. Determinar o valor da variável a partir do pesquisador. Digamos que faço um experimento em que o resultado do lançamento de uma moeda (cara ou coroa), determina se um pedido de acesso a informação será feito por um homem ou mulher (gênero), em um determinado tema.

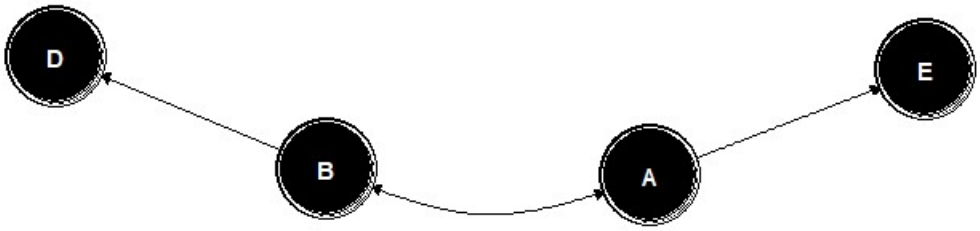


Por outro lado, em um estudo observacional, controlar significa alterar um gráfico original (sem controle). Considere o seguinte gráfico

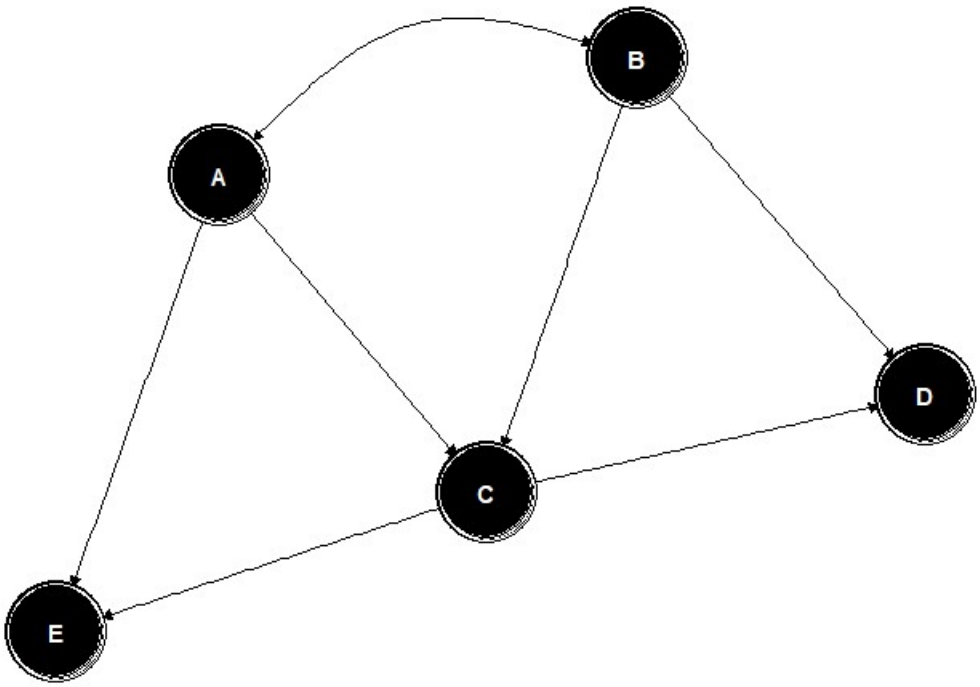


Controlar para C, significa: 1. Eliminar todas as flechas que saem de C. 2. Elimine as flechas que vão para C, se se for um collider, e conecte os parentes de C por meio de linha tracejadas. 3. elimine C





E se eu controlar para G? 1. Eliminar flechas que saem de G. 2. Elminar flechas que saem do collier (no caso, C, que é parente de G) 3. Eliminar G



Em resumo, se C depende de A e B de forma independente, controlar para C significa que, estratificando em C, tenho outra relação entre A e B. Para ilustrar, se A é binária, e B é binária, e C = A + B, entÃo para C = 1, A = 0 e B = 1 ou B = 1 e A

= 0. Logo, se sei informação sobre A, automaticamente determino B e vice-versa.

De modo geral, condicionar em um collider troca o status dos caminhos que passam por C. Caminhos que estavam abertos são fechados, e caminhos que estavam fechados são abertos. E condicionar em um descendente de C também gera esse problema. Caminhos que estavam abertos continuam abertos, mas com efeitos atenuados. E caminhos fechados ficam abertos (com efeitos leves).

Referências

Hernán MA, Robins JM (2019). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming. Disponível (temporariamente) em: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

MixTape

Greenland, S., & Pearl, J. (2014). Causal diagrams. Wiley StatsRef: Statistics Reference Online, 1-10.