Câncer de pulmão e tabagismo

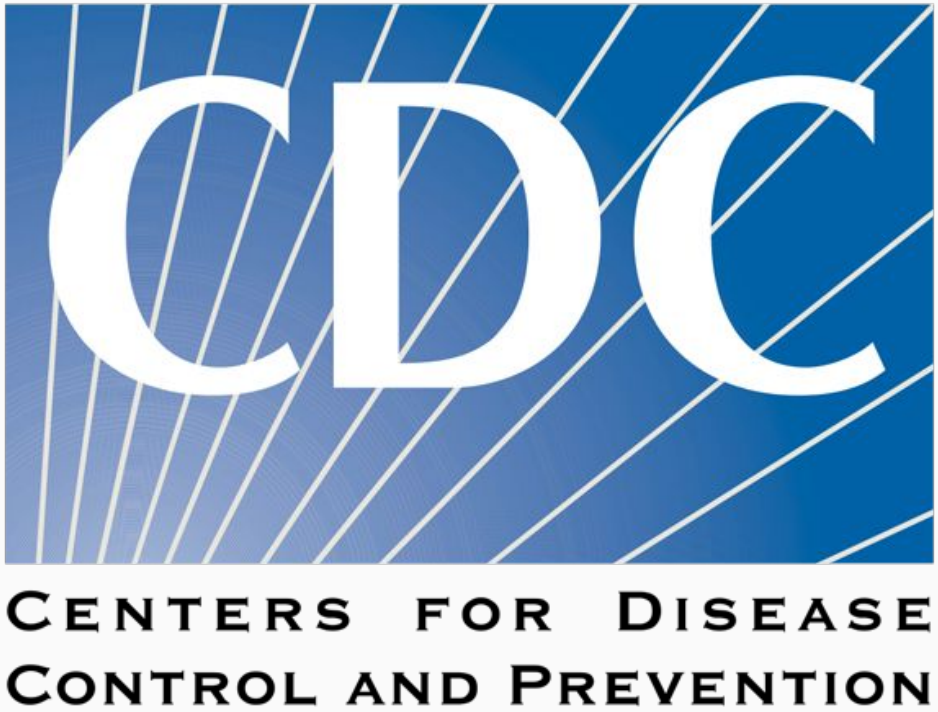
Vitor Horodynski, RA 206953

Introdução

Os principais dados usados para o projeto vem da API Wonder e de uma base de dados, ambas fornecidas pelo CDC.

A API é um registro de mortalidade organizado por causa, ano e mês (entre outros) dos EUA e a base de dados é uma pesquisa que revela os hábitos de consumo de tabaco da população..

Os dados foram filtrados por ano e estado norte-americano.



Extração e Conversão de Dados

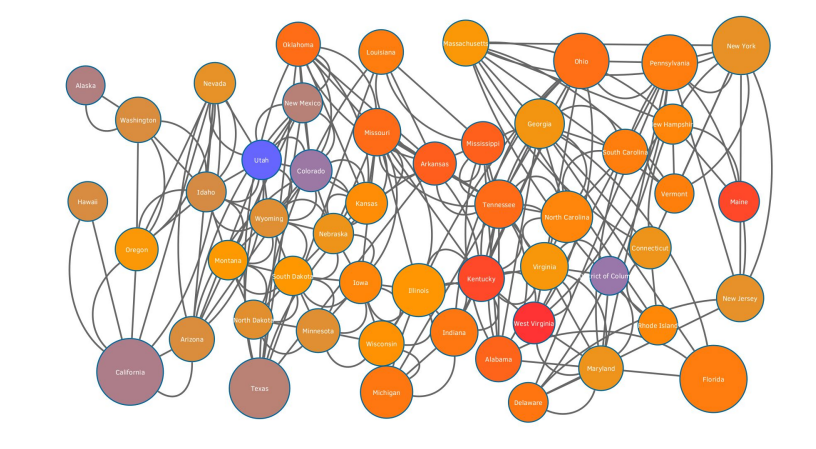
Além dos dados da API Wonder e da base de dados do CDC, foram usados também dados de fontes como o Github e o Kaggle para enriquecer as análises realizadas, tanto de maneira manual pelo download dos dados ou de forma automatizada, no caso da API Wonder e do Github.

A obtenção manual de dados envolvia também o processo de conversão, em que apenas uma parte de uma base é selecionada ou várias bases foram unidas para formar apenas uma que complementava uma série histórica maior. Para tanto, foram usados tanto o Python como o SQL.

Uso dos dados consolidados

A partir do tratamento, conversão e integração de diversas fontes de dados diferentes, o objetivo central foi geral gerar análises visuais tanto da mortalidade por câncer de pulmão nos EUA, o padrão de uso de tabaco e como estas duas métricas se relacionam.

Para tanto foi usado o modelo de grafos, já que este provê uma maneira visualmente natural de observação destes dados.



Mapa de calor da mortalidade por câncer de pulmão nos EUA em 2018

O papel do modelo relacional

Requisições

A consolidação dos dados em SQL foi usada para escrever requisições que tratam da característica dos dados que foram obtidos. Por exemplo:

> Quais são os estados norte-americanos que mais frequentemente ultrapassam a média nacional de mortes por câncer de pulmão a cada ano? Qual o período de tempo com maior número de mortes per capita por câncer de pulmão?

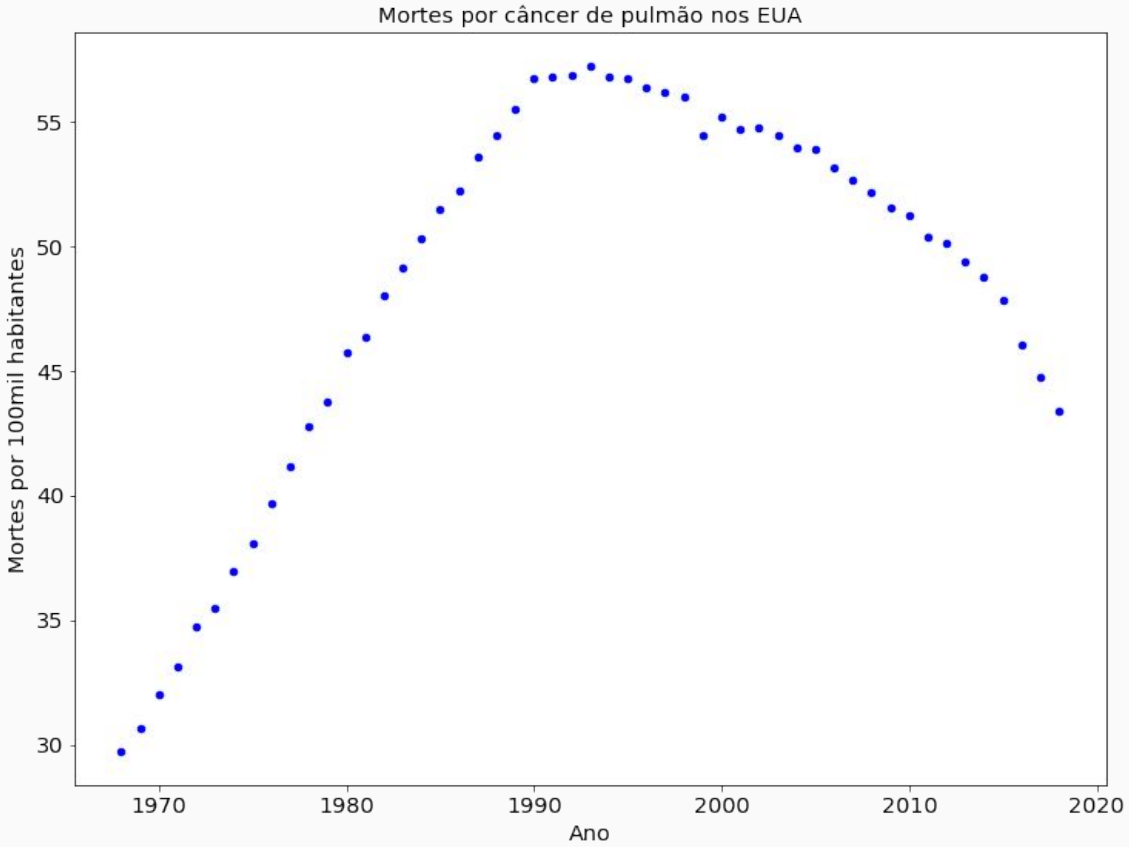
Requisições como estas possibilitam realizar novas descobertas sobre o problema.

70

mortes por câncer de pulmão a cada 100 mil habitantes em média de 1968 a 2018 no estado de West Virginia. Em 1993, foi registrado o máximo valor desta estatística em toda a nação: 57 mortes/100 mil habitantes.

Além disso, o modelo relacional permite construir uma série de artefatos que antes não seriam possíveis sem ele.

O gráfico ao lado, por exemplo, é a consolidação da mortalidade nacional de 1968 a 2018 por câncer de pulmão nos EUA dentre os diversos estados, um dado que não era facilmente disponível antes de sua consolidação.

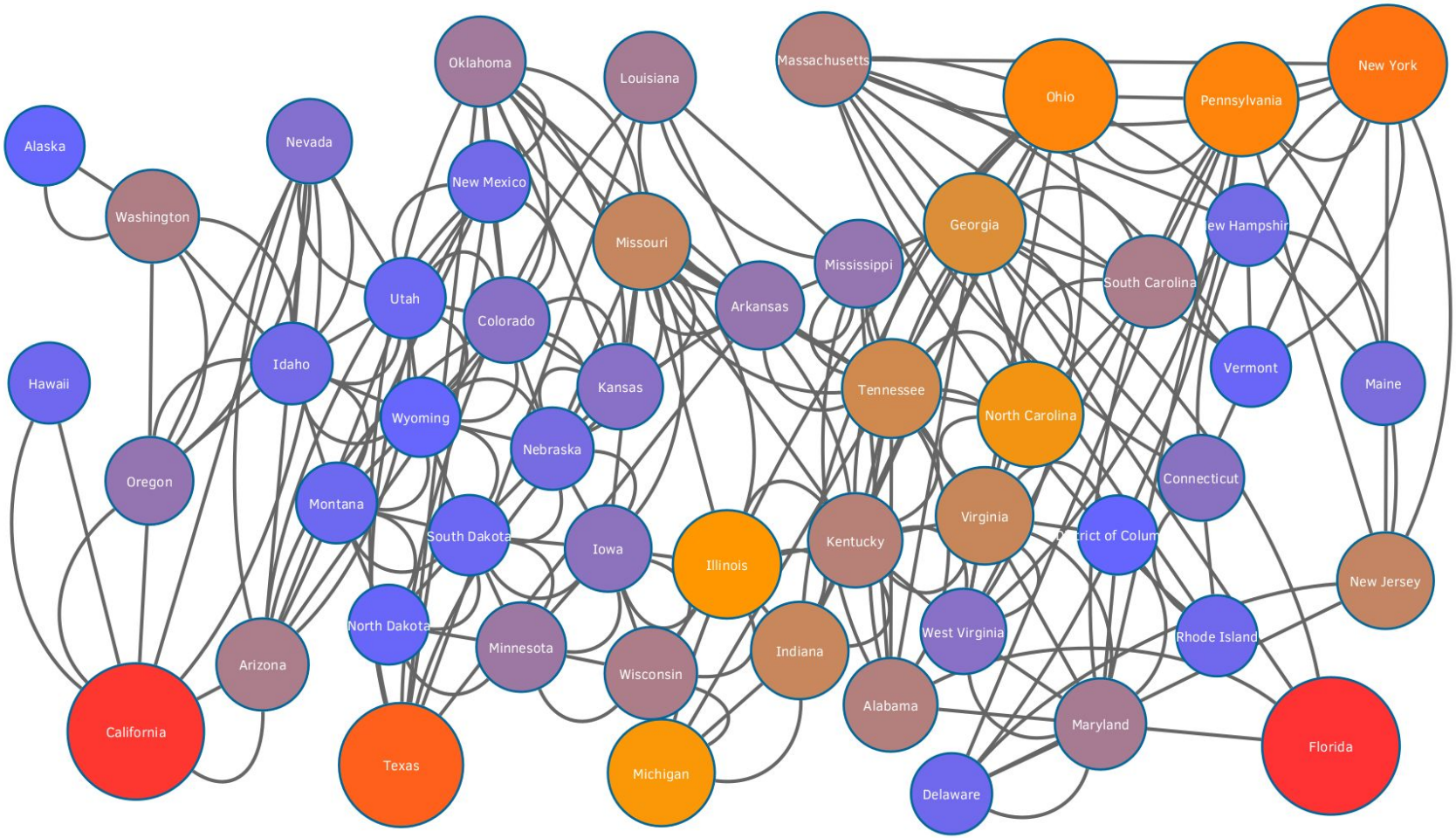


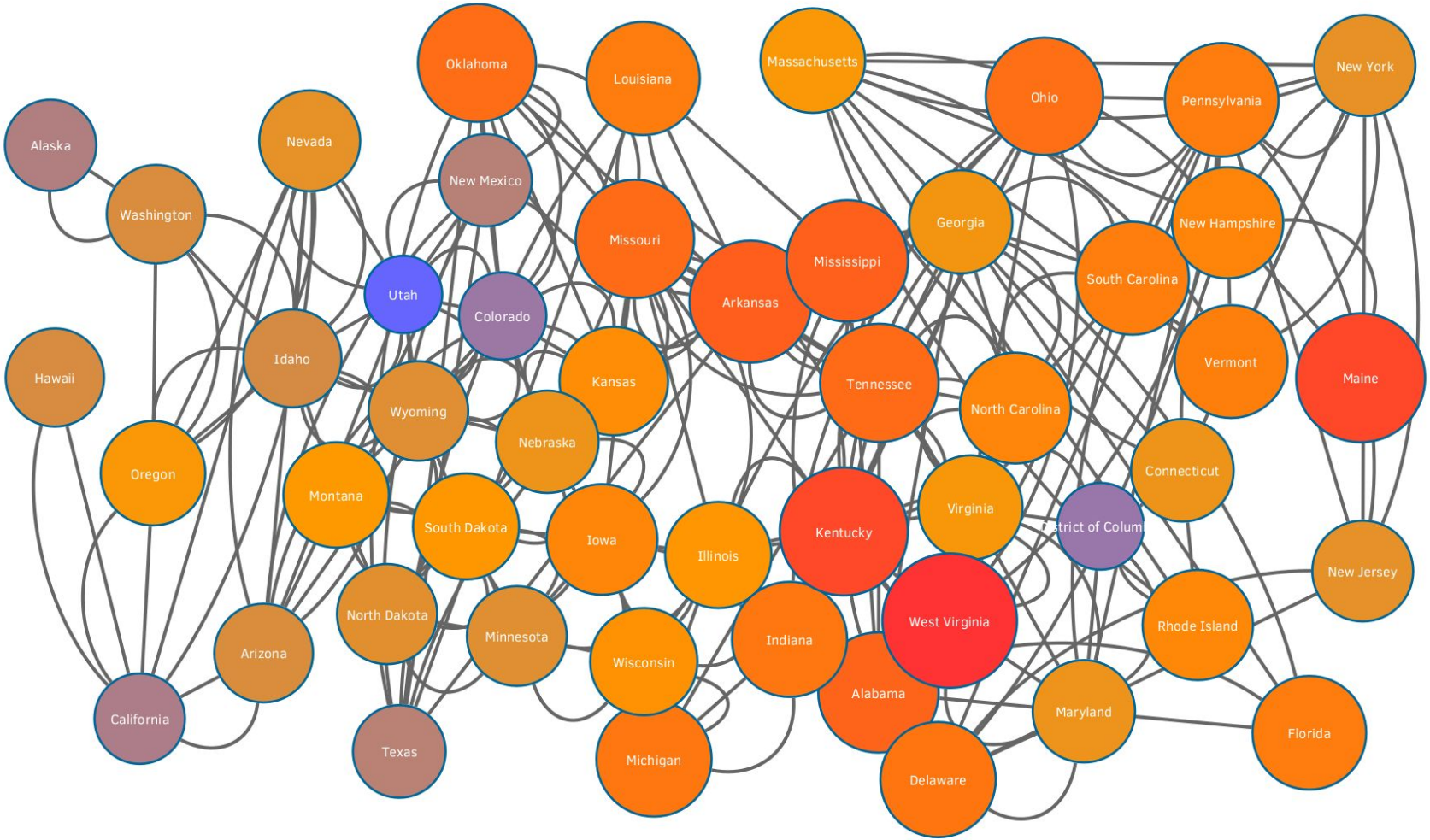
Análises visuais por meio do modelo de grafos

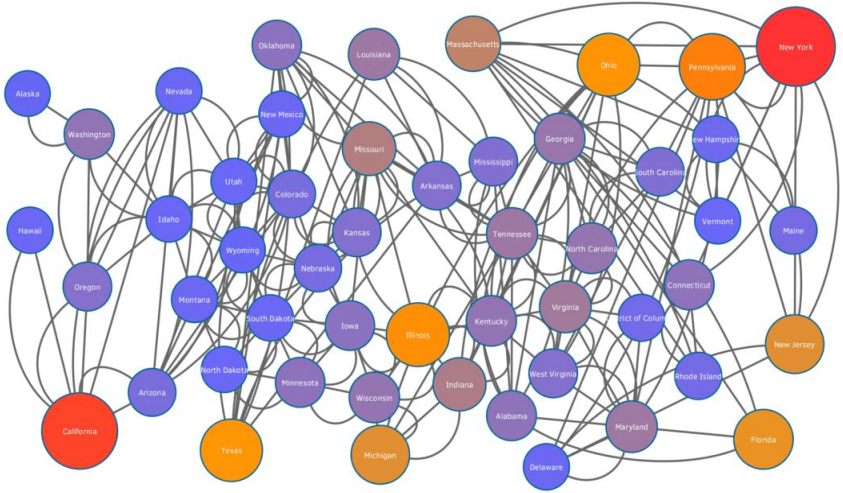
Câncer em números absolutos x Per Capita

Os dois grafos a seguir são um mapa de calor do número de mortes por câncer de pulmão no ano de 2018. Entretanto, um deles foi feito usando o número de mortos em relação à população e outro foi feito com relação ao número absoluto de mortes por estado norte-americano.

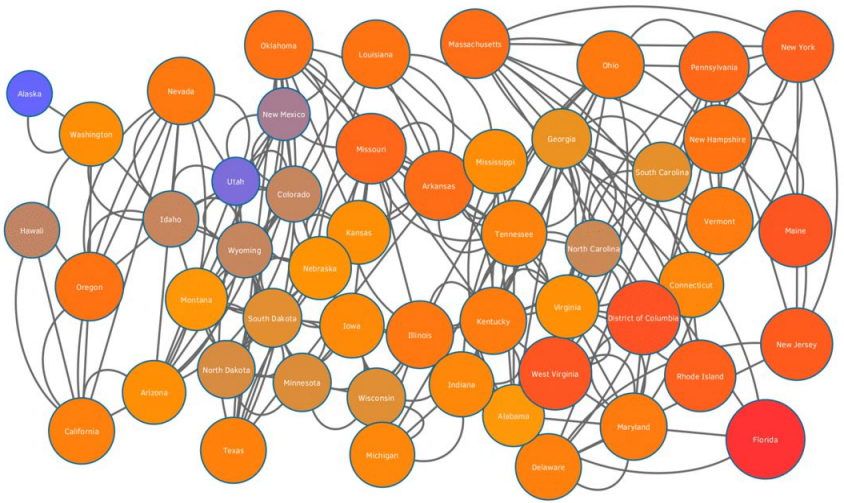
Dependendo de qual contexto de apresentação destes dados, podem ser tomadas decisões completamente diferentes sobre a doença. O já citado estado de West Virginia não chamaria a atenção quando se trata de número total de mortes, apesar de sempre superar a média nacional de mortes por câncer de pulmão.


Oklahoma
Louisiana
Massachusetts
Ohio
Pennsylvania
New York
Alaska
Nevada
New Mexico
Washington
Missouri
Georgia
Mississippi
South Carolina
Utah
Colorado
Arkansas
Idaho
Kansas
Tennessee
Vermont
Maine
Hawaii
Wyoming
North Carolina
Nebraska
Oregon
Connecticut
Montana
Virginia
South Dakota
Iowa
Illinois
Kentucky
New Jersey
North Dakota
Minnesota
Indiana
West Virginia
Rhode Island
Arizona
Wisconsin
Alabama
Maryland
California
Texas
Michigan
Florida
Delaware


Oklahoma
Louisiana
Massachusetts
Ohio
Pennsylvania
New York
Alaska
Nevada
New Mexico
Washington
Missouri
Georgia
New Hampshire
Mississippi
Utah
Colorado
Arkansas
South Carolina
Hawaii
Idaho
Kansas
Tennessee
Vermont
Maine
Wyoming
North Carolina
Nebraska
Oregon
Connecticut
Montana
Virginia
Kentucky
South Dakota
Iowa
Illinois
New Jersey
North Dakota
West Virginia
Rhode Island
Minnesota
Indiana
Arizona
Wisconsin
Alabama
Maryland
California
Florida
Texas
Michigan
Delaware



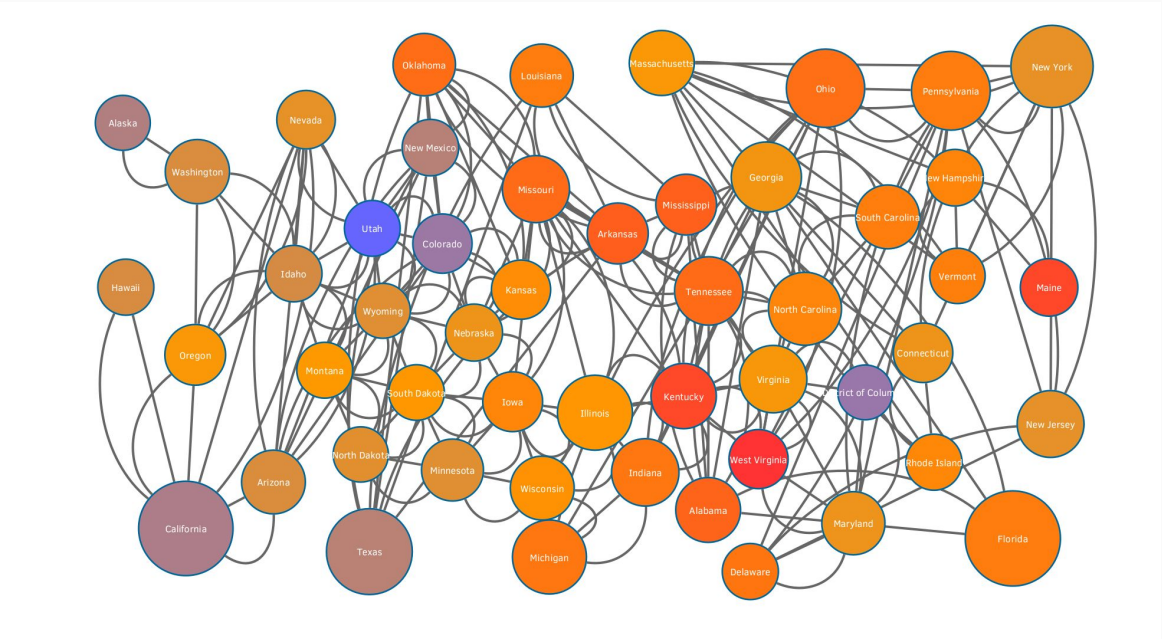
Mapa de calor da mortalidade absoluta por câncer de pulmão nos EUA, de 1968 a 2018



Mapa de calor da mortalidade per capita por câncer de pulmão nos EUA, de 1968 a 2018

Para mitigar esta diferença de interpretação entre estas duas métricas, pode-se usar uma combinação do mapa de calor com o tamanho dos nós que representa cada estado.

No grafo ao lado as cores representam um maior número de mortes (para o vermelho) per capita e o tamanho dos nós é proporcional ao número de mortes totais por estado:

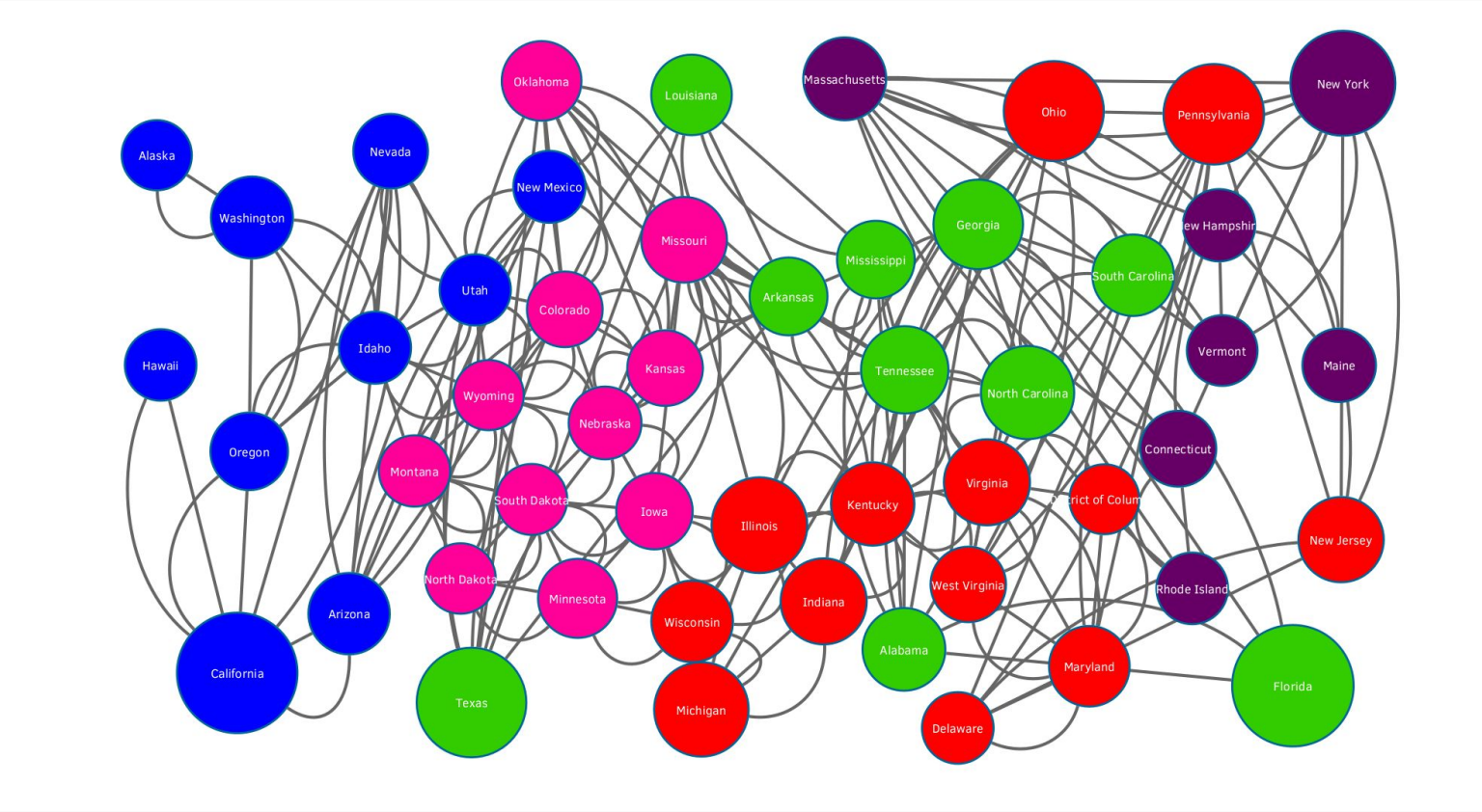


Mapa de calor da mortalidade per capita por câncer de pulmão nos EUA, de 1968 a 2018. O tamanho de cada um dos nós é proporcional ao número total de mortes no estado.

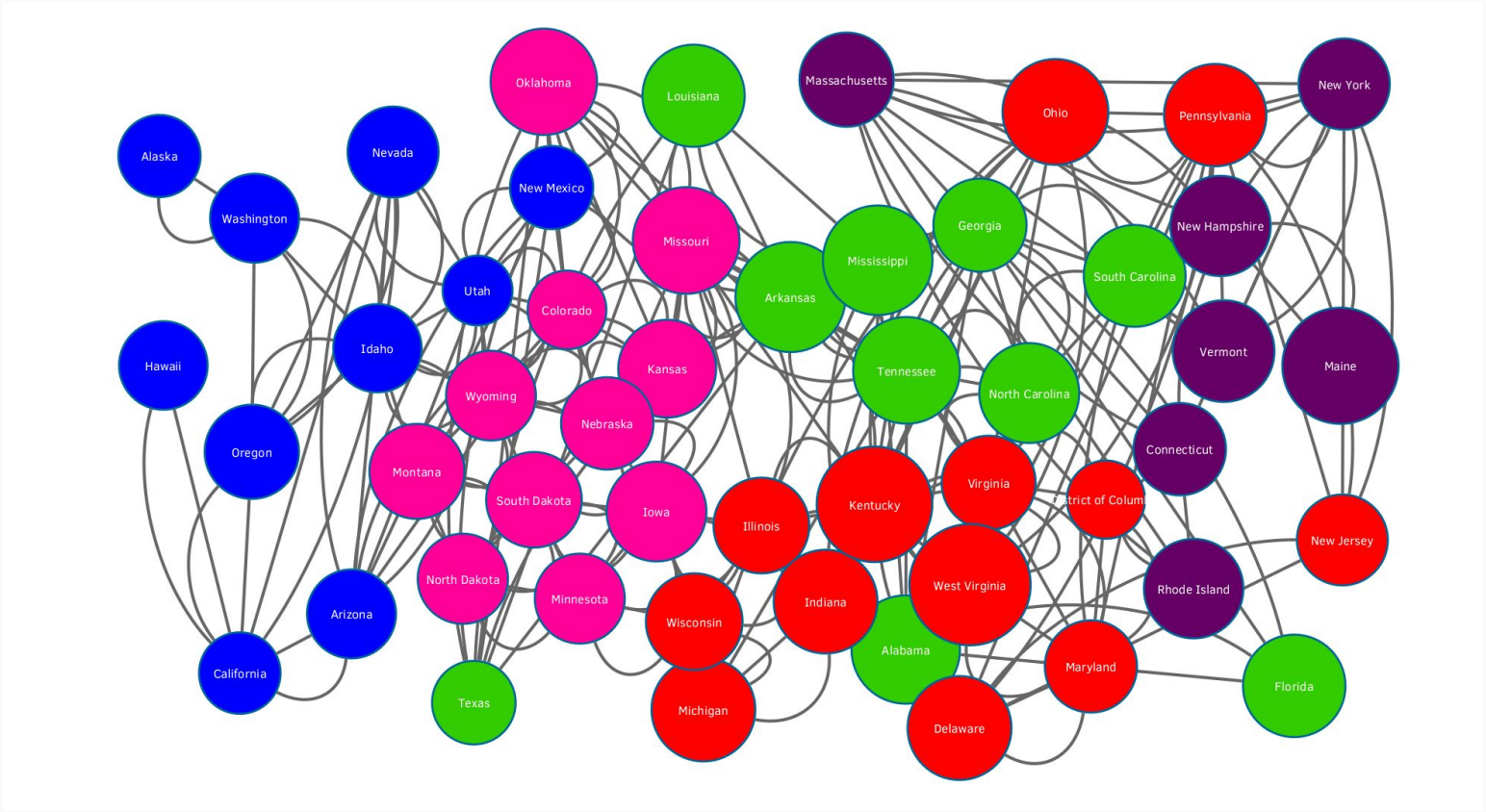
Comunidades no grafo de vizinhanças topológicas

Ao construir o grafo de vizinhança entre estados norte-americanos, uma das perguntas em que pode-se chegar é se existe alguma relação entre os casos de câncer que estão sendo estudados e estas comunidades. Esta hipótese pode ser testada por meio da seguinte análise visual:

Colorir os nós de uma mesma vizinhança com a mesma cor, diferente da cor de outras vizinhanças. Assim, pode-se usar o tamanho do nó para exprimir o número total de mortes ou o número de mortes per capita e então visualizar a conexão (ou não) entre a medida e a comunidade do nó.



Mapa de comunidades das vizinhanças dos estados norte-americanos, onde o tamanho de cada um dos nós representa o número de mortes totais no estado em 2018.

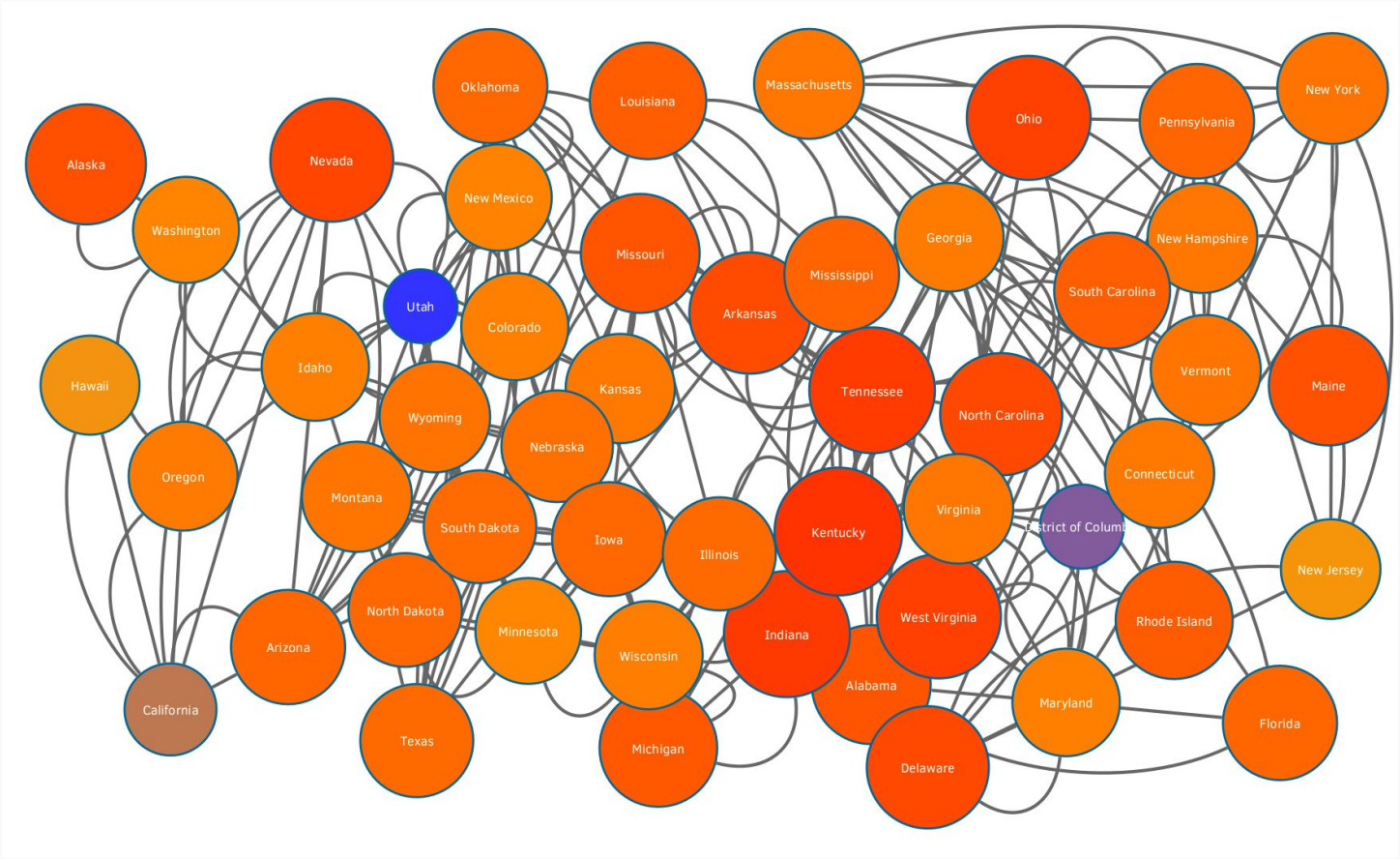


Mapa de comunidades das vizinhanças dos estados norte-americanos, onde o tamanho de cada um dos nós representa o número de mortes per capita no estado em 2018.

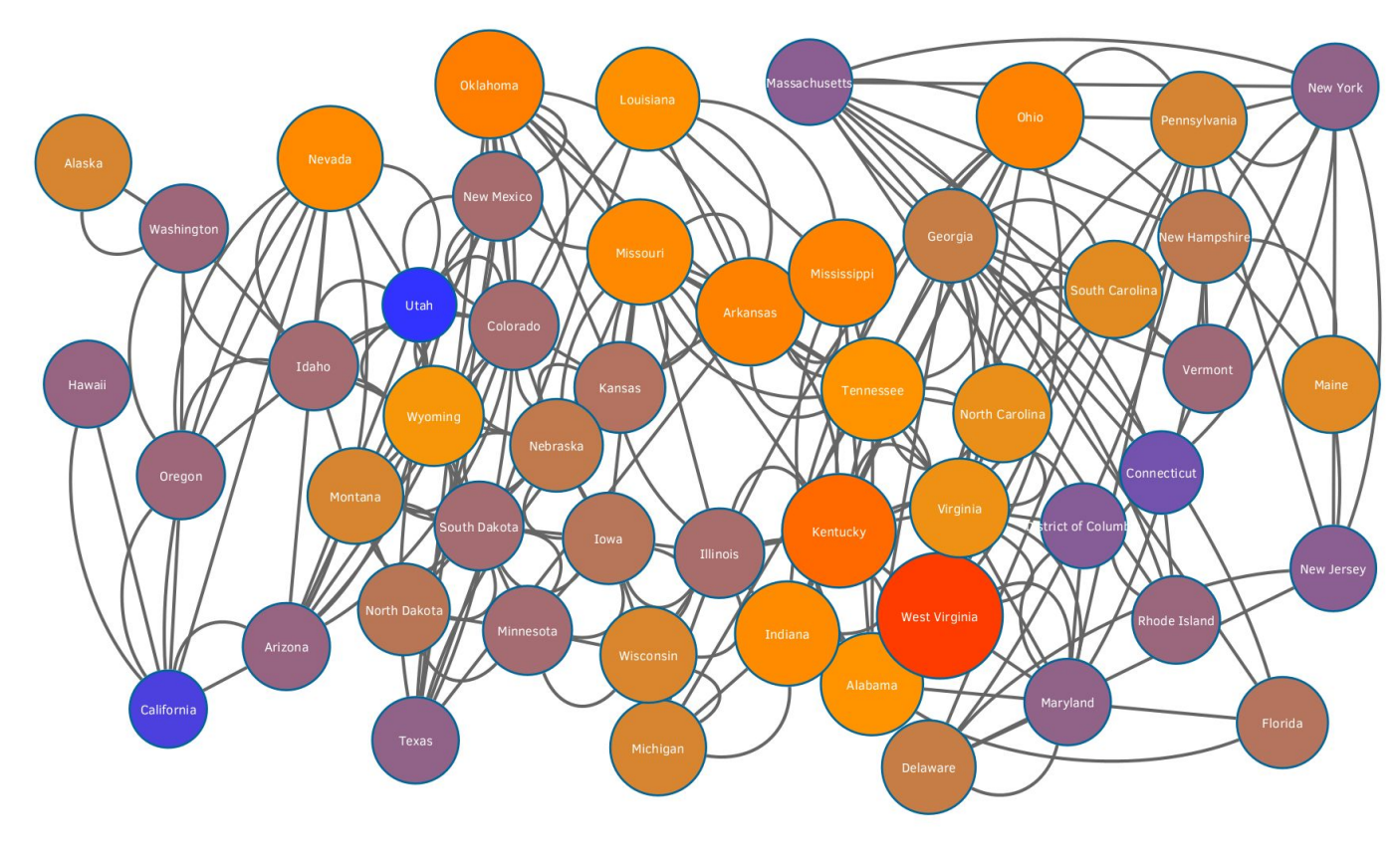
Padrão de consumo do tabaco

A partir da pesquisa sobre o uso do tabaco, é possível construir um mapa de calor dos que declararam fumar todos os dias em 1995 e comparar com o mesmo mapa de calor para o ano de 2010 (último ano em que esta pesquisa esta disponível).

A partir destes mapas de calor e sua comparação visual, entre um período de 15 anos, é possível entender a evolução do padrão de consumo de produtos do tabaco e semelhantes e relacionar esta análise com os recentes movimentos das grandes empresas de tabaco em adquirir empresas de vaporizadores, por exemplo.



Mapa de calor da população dos estados que reportaram fumar todos os dias em 1995.



Mapa de calor da população dos estados que reportaram fumar todos os dias em 2010.

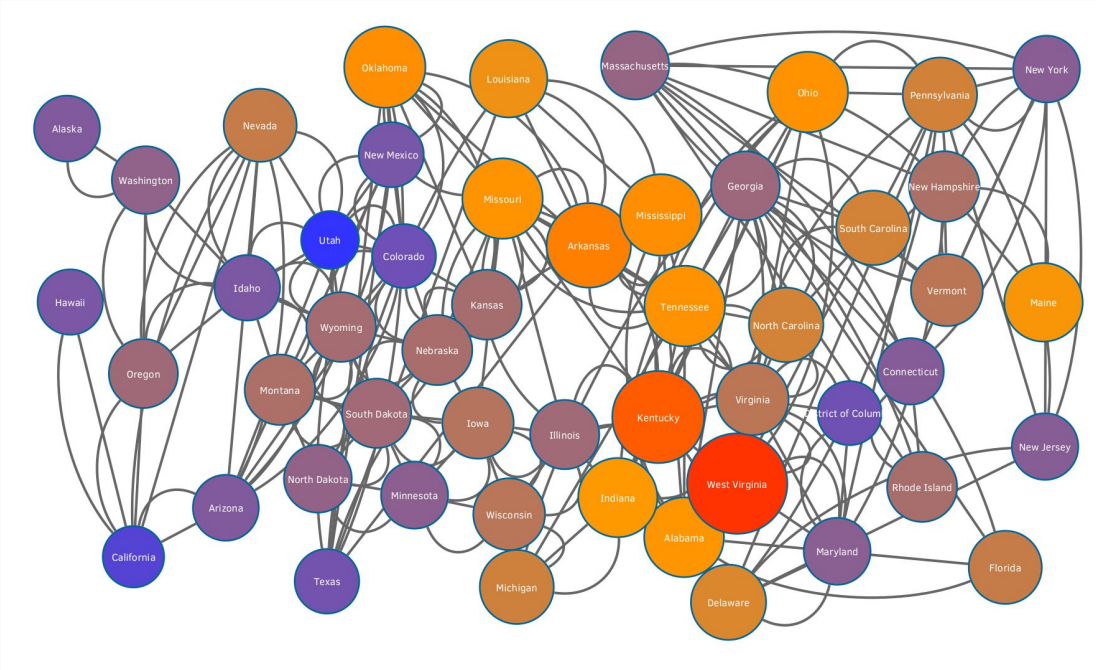
Correlação entre uso de tabaco diário e câncer

Com o objetivo de estabelecer uma correlação entre o uso diário de tabaco e o câncer de pulmão, umas das possíveis alternativas é estabelecer um índice que leva em conta as duas medidas, ambas normalizadas. Assim, desenvolveram-se dois índices:

Multiplicativo: multiplica a porcentagem da população que faz uso tabaco diariamente com o número de mortes por câncer de pulmão per capita.

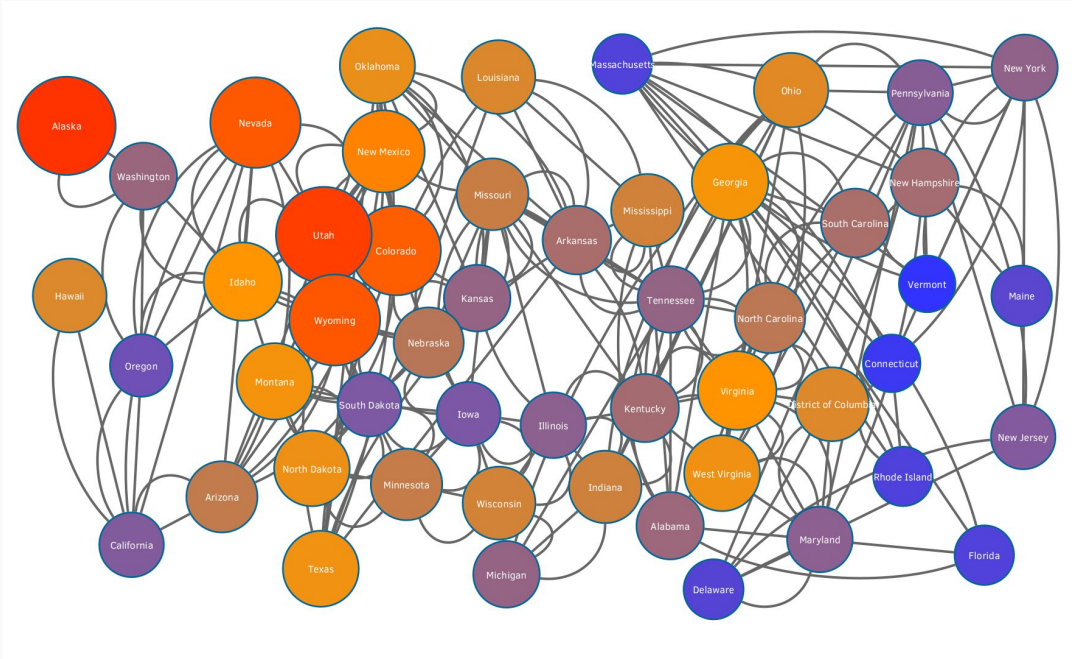
Divisivo: divide a porcentagem da porcentagem da população que faz uso diário de tabaco pelo número de mortes por câncer de pulmão per capita.

É fácil perceber que o índice multiplicativo é falho. Além de ser proporcional a (1/População)^2, se o índice fosse correto, ou seja, fosse aproximadamente constante para qualquer estado, então o número de mortes por câncer de pulmão cresceria com a diminuição do número de fumantes diários, o que conhece-se por não ser verdadeiro.



Mapa de calor do índice multiplicativo em 2010. Nos cálculos deste índice é observada uma enorme variância de índices para cada estado.

Por outro lado, o índice divisivo é adimensional (População/População) e sugere intuitivamente que, se válido, o aumento no número de fumantes diários também está relacionado ao aumento do número de mortes por câncer de pulmão, o que está com o atual conhecimento científico.



Mapa de calor do índice divisivo em 2010. Os cálculos deste índice indicam baixa variância entre cada um dos índices dos estados, o que pode vir da possível correlação entre fumar e ter câncer de pulmão