# Construção e Caracterização de uma Rede para Estudo da Covid-19 nos Municípios Brasileiros

**Cleiton Almeida**[1], **Girolamo Santoro**[1]

[1]Programa de Pós-Graduação em Engenharia de Sistemas e Computação (PESC)
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia (COPPE)
Universidade Federal do Rio de Janeiro (UFRJ)

{cma,santoro}@cos.ufrj.br

***Abstract.*** *This report describes the final work developed in the Complex Networks course. First, we build a network to study COVID-19 dynamic correlations in the brazilian cities. The vertices of the network represents the Brazilian cities; the edges connect those cities with a high correlation degree in their dayly incidence time series. Then we analize some structure properties of the network: degree and distance distributions, clustering, closeness, similarity and assortativity. We find the network is not a scale-free network. The network distance shows a weak but not negligible correlation with the physical distance of the cities. In the same way, the assortativity of the cities regarding region, population and GINI index are low but not negligible. In the other hand, assortativity by GDP is very low. We conclude that the network model is very simple and limited, but still interesting.*

***Resumo.*** *Este relatório descreve o trabalho final do curso de Redes Complexas. Primeiramente, construímos uma rede para estudarmos correlações na dinâmica da COVID-19 nos municípios brasileiros. Os vértices representam os municípios e as arestas conectam cidades que possuem alta correlação em suas séries temporais de incidência diária de casos da doença. Analisamos então algumas propriedades estruturais: distribuição de grau, distância. clusterização, centralidade de closeness, similaridade e assortatividade. Observamos a rede não é livre de escala. Também, as distâncias na rede possuem uma correlação fraca, mas não desprezível, com a distância física dos municípios. Ainda, a assortatividade dos municípios com relação à região, população e IDH é baixa, mas não desprezível. Por outro lado, a assortatividade em relação ao PIB per capita é muito baixo. Concluímos então que o modelo de rede construído, apesar de simples e limitado, ainda assim é interessante.*

## 1. Introdução

### 1.1. Objetivo

O objetivo deste trabalho foi verificar a evolução da doença COVID-19 nos municípios brasileiros e verificar semelhança de comportamento entre municípios vizinhos ou não. Algumas conclusões puderam ser obtidas observando-se a própria estrutura da rede através de métricas específicas para tal, bem como para outras conclusões utilizamos métricas que levaram em conta atributos como por exemplo: PIB per capita, região, IDH, População e distância física.

### 1.2. Revisão bibliográfica

[Qing Cheng 2020] realizaram uma análise em essência bastante similar a este trabalho para algumas cidades chinesas no início da pandemia naquele país. Entretanto, apesar de utilizarem o mesmo tipo de dados (séries temporais de casos de covid) e análise de correlação, os autores não construíram uma rede dos municípios, se limitando às análises estatísticas convencionais.

Ao nosso melhor conhecimento, o tipo de análise desenvolvida neste trabalho, para a epidemia da Covid-19, até então é único na literatura.

## 2. Ambiente computacional

O trabalho foi desenvolvido na linguagem Python 3 com o auxílio das seguintes bibliotecas principais: Networkx, Matplotlib, Pandas, Geopandas Scipy e Numpy. Utlizamos ainda o ambiente de desenvolvimento e testes Google Colaboratory[1]. O Google "Colab" é um ambiente baseado Jupyter Notebook pré-configurado com recursos de computação adequados para protótipos de projetos em *Data Science* e *Machine Learning*. Os notebooks criados neste trabalho estão disponíveis no github de um dos autores [2].

## 3. Construção da rede

### 3.1. *Dataset*

As séries de casos novos de Covid-19 de todos os municípios brasileiros foram obtidas através da página Brasil-IO [3]. A fonte original são os dados disponibilizados pelas secretarias estatuais de saúde. Estes dados foram coletados, processados e "libertados" à sociedade, de forma voluntária, por Álvaro Justen e dezenas de outros colaboradores, aos quais somos gratos pela nobre iniciativa.

Utilizamos também dados de População (IBGE), IDH (IBGE, Firjan), e PIB per Capita (IBGE).

### 3.2. Visão geral

Cada município possui uma curva única de evolução de número de casos novos Covid ao longo dos meses. Esta evolução temporal pode ser interpretada como uma "assinatura" da epidemia em cada município, revelando informações de quando ela iniciou, quando atingiou o pico e qual a tendência ao longo dos meses. Ao compararmos a dinâmica da evolução da Covid em municípios distintos, podemos perceber quais municípios tiveram dinâmica parecida em forma, conforme ilustrado na figura 1.

### 3.3. Pré-processamento

A fim de eliminar algumas inconsistências e diferenças observadas nas séries temporais dos municípios, realizamos as seguintes etapas de pré-processamento nos dados brutos:

- Preenchimento dos primeiros dias da série com número de casos igual a zero, até o primeiro dia com número de casos não nulo;

[1] /https://colab.research.google.com
[2] https://github.com/cleitonmoya/CPS765_Trabalho2
[3] https://brasil.io/dataset/covid19/caso_full/

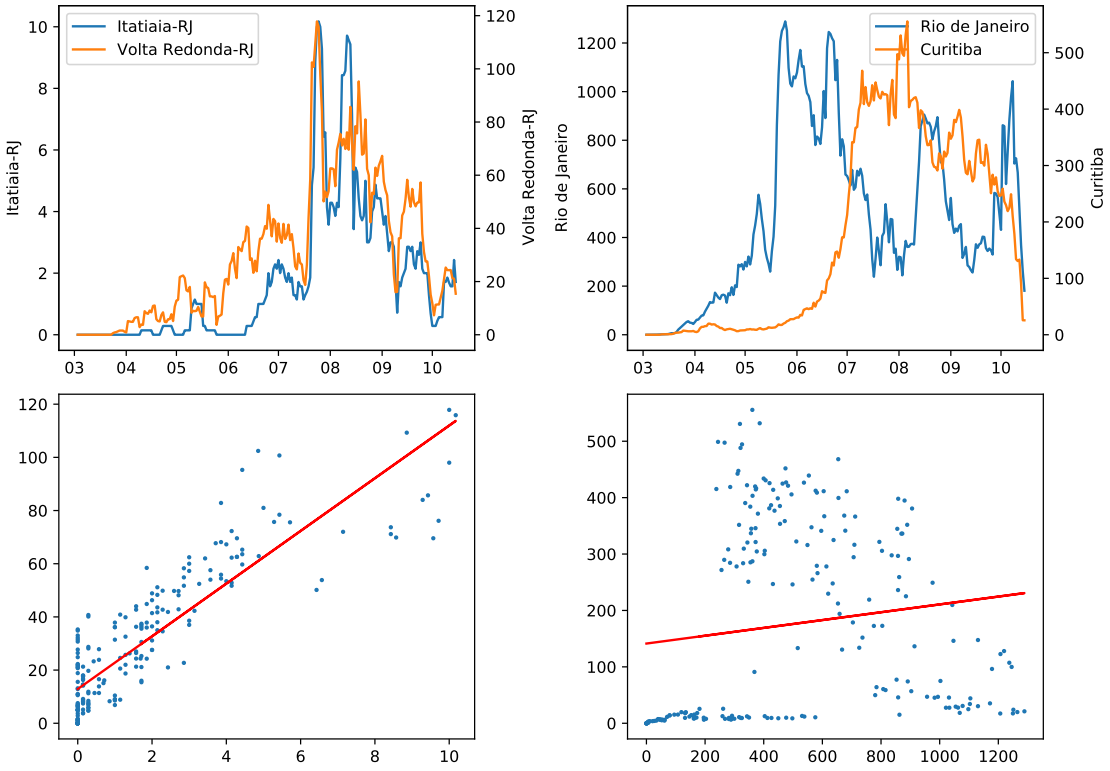


**Figura 1. Dinâmica temporal da Covid-19 em municípios diferentes**

- Interpolação polinomial de segunda ordem para os dias com dados faltantes;
- Aplicação de função para garantir a monotonicidade crescente no número de casos acumulados (ou seja, eliminar os números de casos negativos).
- Aplicação de média móvel de 7 dias para atenuar a alta oscilação dos números causada inclusive por fatores artificiais (como o represamento dos dados nos finais de semana).
- Limitamos o intervalo de análise de 1 de março até 15 de outubro de 2020, período em que tínhamos (na data de início deste trabalho) dados para todos os municípios.

### 3.4. Matriz de covariância e limiar para arestas

Após o pré-processamento, em posse das séries temporais já filtradas, geramos uma matriz de covariância entre os 5.570 municípios.

Para avaliarmos a semelhança da dinâmica de casos novos entre os municípios e, desta forma, estabelecermos os relacionamentos (arestas) utilizamos a correlação de Pearson, gerando assim uma rede baseada em correlação. Esta técnica é bastante utilizada para a construção de redes de diversos domínios, como por exemplo redes biológicas e redes financeiras [Sadamori Kojaku 2019].

A principal vantagem desta técnica de construção é a simplicidade. A rede é obtida diretamente à partir da matriz de correlação, estabelecendo-e um valor limiar. Por outro lado, o maior problema é como selecionar o limiar para filtrar a matriz de correlação de modo a preservar a maior quantidade possível de informação e ao mesmo tempo evitar conexões espúrias. Diferentes abordagens vem sendo proposta na literatura para amenizar

este problema, como [Elisa Benedetti 2020] [Sadamori Kojaku 2019]. Entretanto, neste trabalho escolhemos trabalhar com um *threshold* fixo a fim de verificar se mesmo um modelo simples seria capaz de permitir algumas análises e conclusões interessantes.

Construímos 3 modelos com os seguintes limiares: $0.80$, $0.85$ e $0.90$. Analisamos então as distribuições de grau, a distância média e o número de vértices da maior componente conexa da rede (tabela 1 e figura 2).

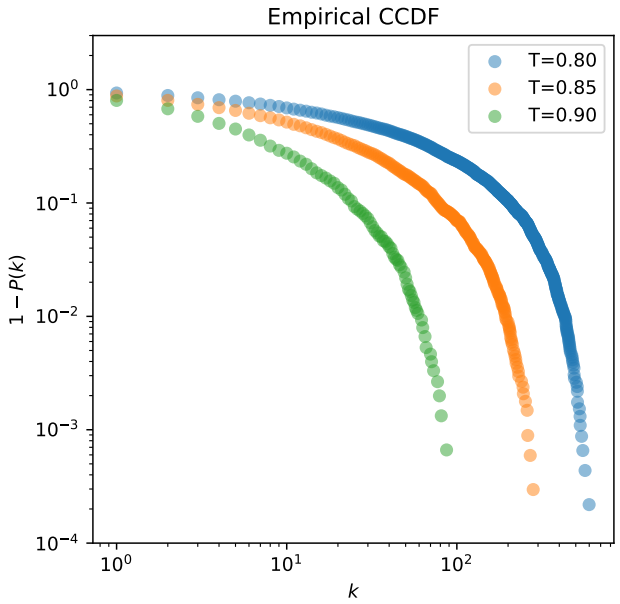


**Figura 2. Distribuição de grau - diferentes limiares**

**Tabela 1. Impacto do limiar na maior componente conexa da rede**

| | $T = 0.80$ | $T = 0.85$ | $T = 0.90$ |
|---|---|---|---|
| Grau máximo: | 608 | 293 | 89 |
| Grau médio: | 70 | 28 | 10 |
| Distância média: | 4 | 5 | 6.4 |
| Tamanho da GCC: | 4575 | 3370 | 1507 |

Na figura 2, observamos que, quanto maior o grau, menos pesada é a cauda. Já na tabela 1, podemos ver que a distância média aumenta com o aumento do limiar. De forma contrária, a maior componente conexa (GCC) diminui com o aumento do limiar. Com base nos valores, selecionamos para análise neste trabalho o limiar $T = 0.85$. No restante do trabalho, realizamos todas as análises utilizando somente a GCC.

A figura 3 mostra a rede construída (maior componente conexa incluindo os vértices das capitais em destaque).

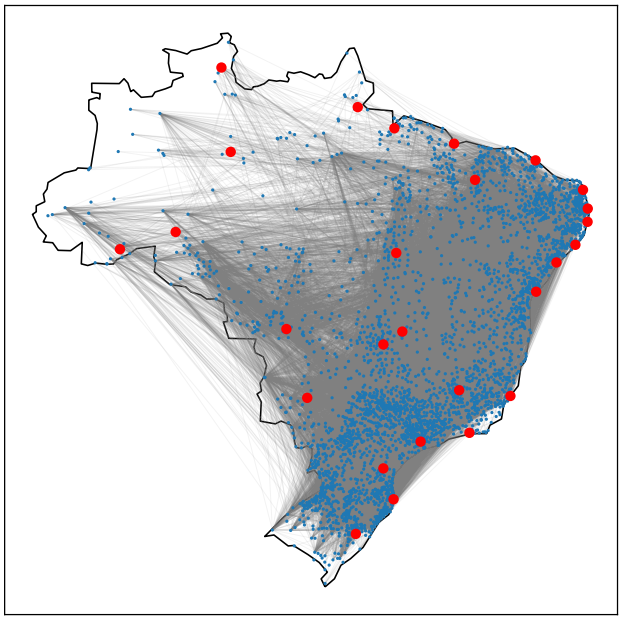

**Figura 3. Rede de correlação da dinâmica da Covid nos municípios brasileiros**

## 4. Caracterização e análises

### 4.1. Distribuição de grau

Como podemos ver na figura 2, a distribuição não é livre de escala (não segue uma lei de potência). Porém a cauda é (relativamente) pesada. A tabela 2 apresenta as estatísticas básicas da distribuição de grau.

**Tabela 2. Distribuição de grau - estatísticas básicas**

| | |
|---|---|
| Máximo: | 293 |
| Mínimo: | 1 |
| Médio: | 28 |
| Mediana: | 11 |

### 4.2. Distribuição de distâncias

A figura 3 mostra a distribuição de distâncias (saltos) na rede. Observamos que distância média na rede é 5 e o diâmetro da rede 18.

### 4.3. Distâncias físicas x distâncias na rede

Uma questão que surge é se as distâncias físicas (geográficas) entre as cidades possuem alguma relação com a distância na rede. Intuitivamente, esperamos que cidades geograficamente possuam dinâmica da Covid com maior correlação do que aquelas distantes geograficamente.

**Tabela 3. Distribuição de distâncias (PMF)**

| | |
|---|---|
| Máximo: | 293 |
| Mínimo: | 1 |
| Médio: | 28 |
| Mediana: | 11 |

Para estimarmos as distâncias físicas entre os municípios conectados na rede, utilizamos suas coordenadas geográficas (latitude e longitude) e fórmula de Heaversine [4]. Calculamos a distância física para cada par de vértices $(u, v)$ conectados, e correlacionamos com a distância na rede entre estes mesmos pares. Utilizando a correlação de Pearson, o valor encontrado foi de $\mathbf{0.17}$.

Concluímos então que a correlação entre distâncias físicas e distâncias na rede é fraca, porém não desprezível.

### 4.4. Clusterização e Centralidade de Closeness

A clusterização local mede a probabilidade de dois vizinhos de um vértice também serem vizinhos. Já a centralidade de *closeness* é definida em função da distância média do vértice com o restante da rede: $C_v = \frac{\sum_{t \in V-v} d(v,t)}{n-1}$ [Figueiredo 2015].

Na figura 4, podemos observar que as cidades com maior clusterização tendem ser aquelas posicionadas em regiões mais densas de cidades. Já as cidades com maior *closeness*, menos correlacionadas (na média) com o restante das outras, tendem a ser aquelas afastadas do litoral e de regiões densas.

Convém comentar que alguns autores e pacotes computacionais (como o networkX) utilizam a definição inversa de *closeness* daquela definida em [Figueiredo 2015], a qual utilizamos na análise e na figura 4. Na definição inversa, cidades com maior *closeness* são aquelas que tendem a ter menor desvio (na média) com a correlação das outras cidades da rede. Ou seja, são as cidades que melhor representam na média a dinâmica da Covid-19 no Brasil.

### 4.5. Similaridade

Realizamos ainda uma análise da métrica de Similaridade de Adamic Adar, definida para cada vértice pela soma ponderada inversamente pelo grau dos vizinhos em comum [Figueiredo 2015]. Inicialmente, esperávamos que cidades semelhantes possuíssem em geral maior correlação, e cidades com baixo coeficiente menor correlação. Entretanto, os resultados não foram aderentes a esta hipótese.

### 4.6. Assortatividade

O coeficiente de assortatividade é uma métrica que mede a intensidade de relacionamento entre vértices com os mesmos atributos e vértices com atributos diferentes. Para um coeficiente $r = 0$, os relacionamentos são aleatórios. Quando $r = 1$, os relacionamentos são somente entre vértices iguais. Já $r < 0$ significa que os relacionamentos são entre vértices diferentes [Figueiredo 2015].

[4] `https://en.wikipedia.org/wiki/Haversine_formula`

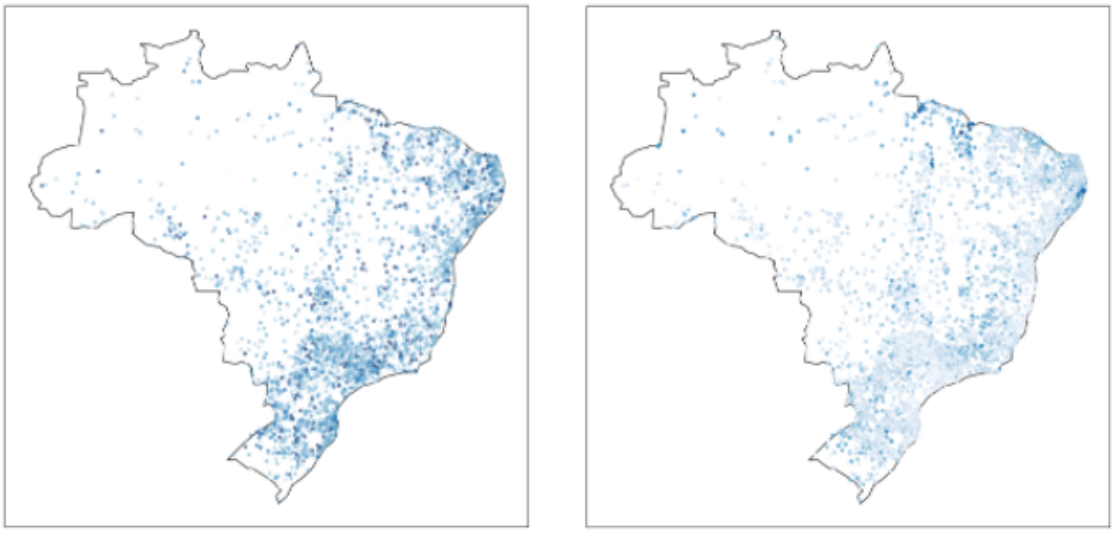


**Figura 4. Clusterização local e centralidade de *closeness***

Inicialmente analisamos a assortatividade de grau. O valor encontrado $(0.32)$ mostra que cidades com grau elevado (*hubs*) tendem a se conectar.

Analismos também a assortatividade para os atributos "região", "faixa populacional", "faixa IDH" e "faixa de PIB *per capita*". Os coeficientes de assortatividade calculados são mostrados na tabela 4. A figura 5 mostra a mixagem por faixa de IDH e PIB *per capita*.

**Tabela 4. Coeficiente de assortatividade - Grau e atributos**

| | |
|---|---|
| Grau | 0.32 |
| Região | 0.17 |
| Faixa populacional | 0.26 |
| Faixa IDH | 0.18 |
| Faixa PIB *per capita* | 0.08 |

É interessante observar na figura 5 que, apesar da maioria das cidades possuírem IDH médio, a maior assortatividade ocorre para cidades com IDH alto. Também, a assortatividade entre cidades com baixo IDH é menor do que cidades com IDH muito alto, o que mostra que cidades com IDH alto possuem um comportamento mais "homogêneo" com relação à dinâmica da epidemia.

Um fator que talvez tenha influenciado nos valores de baixa assortatividade é o fato de não termos eliminado a escala temporal das análises. Por exemplo, cidades com mesmo IDH ou faixa mesma populacional, porém nas quais a epidemia começou em período diferente, em nosso modelo não estão conectadas (baixa correlação de Pearson da série temporal).

## 5. Conclusões

Neste trabalho construímos e analisamos uma rede de correlações da dinâmica do Covid-19 dos 5.570 municípios brasileiros. A rede foi construída a partir da matriz de correlação

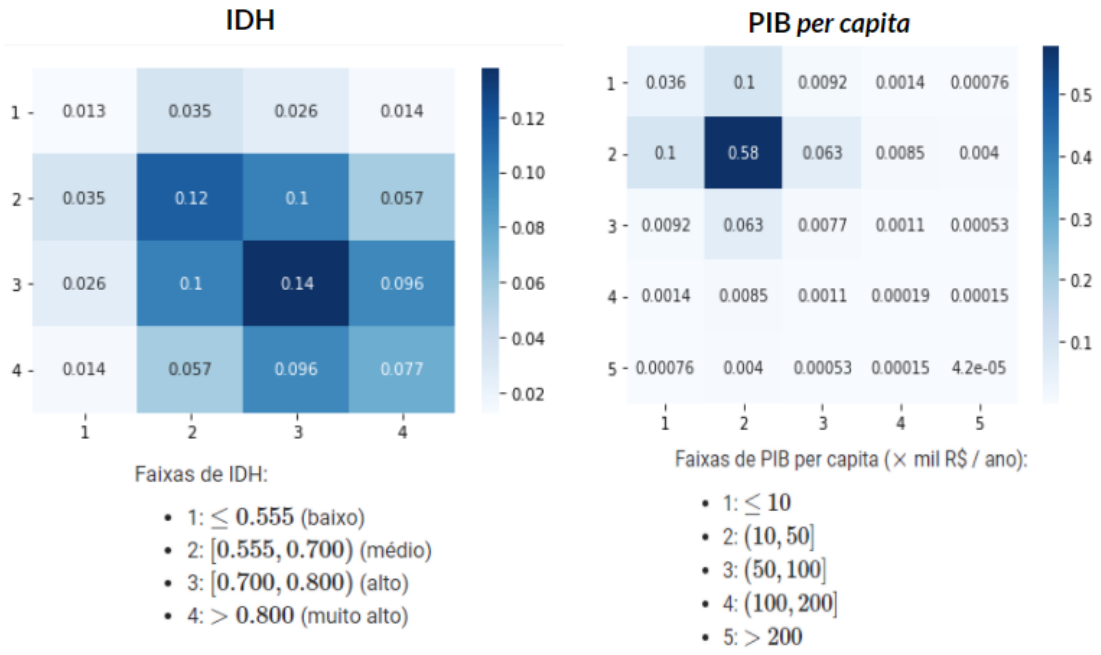


**Figura 5. Mixagem por IDH e PIB *per capita***

da serie temporal de casos novos (média móvel de 7 dias) filtrada a partir de um limiar $T = 0.80$.

Na caracterização, observamos que a rede construída não é livre de escala, apesar de possuir cauda relativamente pesada. Observamos também que a distância média, para o limiar adotado, foi de $5$ saltos e que a distância na rede possui correlação, ainda que fraca, com a distância física entre os municípios. Também observamos que a rede possui assortatividade relativamente alta em relação ao grau $(0.32)$, porém assortatividade baixa (ainda que não desprezível) em relação à região, faixa populacional e IDH. Por outro lado, a assortatividade em relação ao PIB per capita é muito baixo.

Apesar do modelo de rede construído ser bastante simples e limitado, concluímos que ainda assim ele foi capaz de fornecer algumas análises interessantes.

## Referências

Elisa Benedetti, Maja Pučić-Baković, T. K. N. G. M. B. T. t. M. H. J. S. I. R. O. P. C. H. H. A.-A. K. S. G. K. G. L. J. K. (2020). A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nat Commun 11*, (5153).

Figueiredo, D. R. (2015). Redes complexas - aulas 3, 4 e 5 (slides do curso).

Qing Cheng, Zeyi Liu, G. C. J. H. (2020). Heterogeneity and effectiveness analysis of covid-19 prevention and control in major cities in china through time-varying reproduction number estimation. *Sci Rep*, (10).

Sadamori Kojaku, N. M. (2019). Constructing networks by filtering correlation matrices: a null model approach. *Proc. R. Soc. A*, (475).