**COVID-19 Transmission Network: Analyzing of GISAID Network Using Real Stories**

**CSc 8550 Advanced Algorithms – Final Project Report**

**Mokter Hossain, Javad Rafiei Asl, and Zafar Iqbal**

**Georgia State University**

**Abstract**

Since the outbreak of the novel coronavirus pandemic many media, agencies, and organizations spent their time and effort to present the COVID-19 infection stories and transmission networks. However, we found a lack of research that analyzes and verified the authenticity of the infection stories and transmission networks. GISAID is one of the popular organizations that provides the most complete collection of genetic sequence data of influenza viruses and related clinical and epidemiological data the WHO Collaborating Center through its own database [1]. It also provides public access of the genomic epidemiology of COVID-19 data through beautiful transmission network graphs [2]. In this study we developed our own network graph, that we call the benchmark graph, by analyzing about hundred instances of the COVID-19 real infection and transmission stories. Then we analyzed, compared, and contrasted the GISAID presented genomic transmission network graphs with our benchmark graph. This study also intended to find the stories behind how the Novel COVID-19 Transmission Outbreak occurred during the time period December 2019 to April 2020.

**Introduction**

Since mid-December of 2019, the unknown acute respiratory tract infection of coronavirus broke out first in Wuhan, the capital of Hubei province, in central China [3, 11]. Although on December 29, 2019 China reported its first 4 COVID-19 cases. Those initial coronavirus infected people were linked to the Huanan Seafood Wholesale Market in Wuhan, China [3, 4, 5, 6, 11]. Initially, a cluster of patients were presented with an unidentified form of viral pneumonia, caused by a newly identified β-coronavirus, who had common history of visiting the Huanan Seafood Wholesale Market in Wuhan. Soon it rapidly spread across China and many other countries. On **January 23, 2020 China imposed strict lockdown in Wuhan. However,** by 29 January, the virus spread to all provinces of mainland China. All provinces of mainland China initiated the highest response level to public health emergency [11, 14, 15, 16]. WHO declared the outbreak a "Public Health Emergency of International Concern" on 31 January 2020 for fear that the virus spread beyond China [12, 13, 14]. **On March 13, 2020, US president Donald Trump declares national emergency.** As of April 30, 2020, COVID-19 has affected more than 3.5 million patients with about 230 thousand deaths in about 200 countries/regions in the world. Thus, the COVID-19 became a major global health concern.

The global impact of the new COVID-19 epidemic is yet uncertain. However, there is no doubt that has collapsed the entire world. The coronavirus was initially named as the 2019-novel coronavirus (2019-nCoV) on 12 January 2020 by World Health Organization (WHO). Later on**, February 11, 2020 the WHO renamed the novel coronavirus as COVID-19 [1**4, 15, 16**].**

GISAID provides data that could be used to get a transmission network of the COVID-19 disease. Our aim is to collect our own data using scientific journals, credible newspapers outlets and other sources and use this newly compiled dataset to validate the transmission network based on the processing of genetic code of the 2019 coronavirus.

**Literature Review**

Analysis of genome sequence data from pathogens is a novel and useful way to investigate the transmission pattern of a virus. Such kind of analysis is based on the genetic mutation pattern of Genome sequence of different strains of the virus. The objective is to recover the epidemic transmission tree, which points out who affected whom in the transmission network.

The similarity between two sequences has an intuitive meaning. If both the sequences are close enough, it means they have a common ancestor which could be extended further to conclude that the hosts were very close to each other in the transmission tree. Furthermore, if there are not very different from their ancestor then both the hosts should be close to its ancestor in the tree. This kind of logical reasoning opens the possibility to construct a tree from the genome sequence data.

Coronaviruses have crown like spikes on their surface which is why they were named as such. There are seven different sub-groupings of this virus. Although there are many coronaviruses, but this grouping is based on the virus’s tendency to transmit to human beings.

* 229E (Alpha coronavirus)
* NL63 (alpha coronavirus)
* OC43 (beta coronavirus)
* HKU1 (beta coronavirus)
* MERS-CoV (beta coronavirus causes Middle east respiratory syndrome)
* SARS-CoV (beta coronavirus causes severe acute respiratory syndrome)
* SARS-C0V-2 (beta coronavirus caused COVID-19)

Since the first reported case of COVID-19 from Wuhan, China, there has been considerable debate on the origin of this virus that has taken the world by storm. According to worldometers.info, as of April 30, there has been 3,284,922 reported cases and 232,326 deaths.

Based on many studies it seems improbable that SARS-CoV-2 was prepared in laboratory using genetic manipulation. Instead, a plausible explanation would be that it transmitted from an animal host to human. Given the similarity between SARS-Cov-2 and SARS-CoV, it is possible that bats serve as hosts. Malyan Pangolins also possess virus similar to SARS-CoV-2. Therefore, any of these two could be the initial host of the virus. It is possible that a progenitor of SARS-CoV-2 jumped into humans and there is strong possibility that there was some intermediate animal host or hosts before jumping to humans.

A virus spreads by replicating itself many times within the host. This frequent replication could result in copying mistakes which results in small changes in the genetic code of the virus. This process of slight changes in the genetic sequence is called mutation. Figure 1 shows how virus mutation occurs. In case of the novel coronavirus, it mutates twice a month. Coronaviruses, in general, have a genome made of RNA. RNA based viruses are highly conserved and during mutation, very few changes occur. This property of RNA makes it very useful for measuring the evolutionary distance and relatedness of one RNA virus to another. Usually , RNA viruses have a high mutation rate. In contrast to this general behavior, both SARS-CoV AND MERS-CoV have low mutation rates. They have a very low potential of sustained community transmission. On the other hand, SARS-CoV-2 has been proved more lethal then its predecessor owing to its high potential of community transmission.

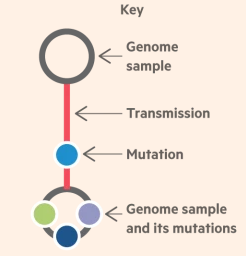


Figure 1: virus mutation

**Transmission Stories of COVID-19**

On 31 December 2019, Wuhan CDC admitted that there was a cluster of unknown pneumonia cases related to Huanan Seafood Market in Wuhan, China. Soon the COVID-19 disease was transmitted from Wuhan to cities outside mainland China via air travel between 25 December 2019 and 19 January 2020 [3,13].

**On January 11, 2020 China reported its first coronavirus death**, a 61-year-old man who had purchased goods from the seafood market **[**aljazeera.com**].** By January 13, 2002 his five other relatives lived in Anyang, China were infected by one of their relatives of lived and got infected in Wuhan [3, 14].

On January 13, 2020, Thailand confirmed its first outbreak of the coronavirus disease that was the first case of coronavirus infection outside China. The patient was with some incoming travelers as either visitors or residents returning from China [19]. On 16 January 2020, COVID-19 was first confirmed to have spread to Japan in a person who traveled to Wuhan [20].

On January 19, 2020, COVID-19 was confirmed in United States by a 35-year-old man who returned to Kirkland, Washington on January 15 after traveling to visit family in Wuhan [21]. Later, the first COVID-19 death in United States was reported on February 28, 2020 after a man with no travel history to China died at in Kirkland in Washington state [22].

On 20 January 2020 South Korea reported the first case of novel coronavirus. The case is a 35-year-old female, Chinese national, residing in Wuhan [23]. On 24 January 2020 the first COVID-19 case in France through five confirmed cases were the individuals recently arrived or returned from Wuhan [24].

The COVID-19 disease first arrived in Canada on January 25, 2020, after a man returned to Toronto from travel in China, including Wuhan [25]. It was also first detected in Malaysia on 25 January on travelers from China arriving via Singapore [26]. It was confirmed to have reached in Victoria, Australia when a man returning from Wuhan [27].

The first COVID-19 case in California, United States was confirmed on January 26. In Orange County, a man in his 50s who was diagnosed with coronavirus after traveling to Wuhan. Later, on February 1, he was released from the hospital [28].

The first COVID-19 case was confirmed in Germany on January 27, 2020. The majority of the cases in January and early February originated from the headquarters of a car parts manufacturer there [29]. It was announced in the United Arab Emirates on 29 January 2020. The first patient, a 73-year-old Chinese woman, was released on 9 February after recovering [30].

On 30 January, India confirmed its first COVID-19 case in a student who had returned from Wuhan University to Kerala. In early February, two other cases were confirmed in Kerala in people who had also been in China [31]. All three of them were successfully recovered.

The first confirmed COVID-19 case to have spread to Italy on 30 January 2020, when two Chinese tourists in Rome tested positive for the virus [32]. One week later an Italian man repatriated back to Italy from the city of Wuhan, was hospitalized and confirmed as the third case in Italy [33].

On January 30, the Centers for Disease Control and Prevention (CDC) had confirmed the**first case of person to person transmission in United States through a** man in Chicago, Illinois case who had returned from Wuhan on January 13 and who tested positive for the virus on January 24 [34].

On 31 January, two members of a family of Chinese nationals staying in a hotel in York, United Kingdom, one of whom studied at the University of York, became the first confirmed cases of COVID-19 in the UK [35]. On 31 January 2020, the first cases of COVID-19 were also confirmed in Russia, Spain, and Sweden by people returned from Wuhan [36, 37, 38].

Florida became the seventh state on March 1 to confirm its first COVID-19 cases: one in Manatee County and one in Hillsborough County, who returned from Italy, where there was a large outbreak [39]. Georgia's first two cases of COVID-19 were reported on March 2, 2020 residents of Fulton County, Georgia who lived in the same household and one had just returned from Rome Italy. Georgia’s first COVID-19 death was confirmed on March 12, 2020 [40].

**Dataset Description**

Technological advances have made it possible to rapidly sequence RNA and DNA genomes on massive scale. This kind of data provides a new source of information to infer the paths of infection. Scientists studying mutations in coronavirus have decoded more than 10,000 different genomes of the deadly pathogen. The analysis of such data is helpful in Creating a comprehensive map that will be crucial to controlling the pandemic and developing medicines to treat it.

One such repository of Genome sequences is GISAID, which is a public-private venture ship between the government of Germany and non-profit organizations. This initiative promotes the rapid sharing of data about the influenza viruses such as SARS-Cov-2. The data is freely available for researchers all around the globe. Figure 1 shows a screenshot of genome sequences database of SARS-Cov-2 available at GISAID repository. In figure 2, we can observe a sample of genetic code sequence of one particular strain of the novel coronavirus named as hCoV-19/USA>NY-NYumc291|EPI\_ISL\_428808.

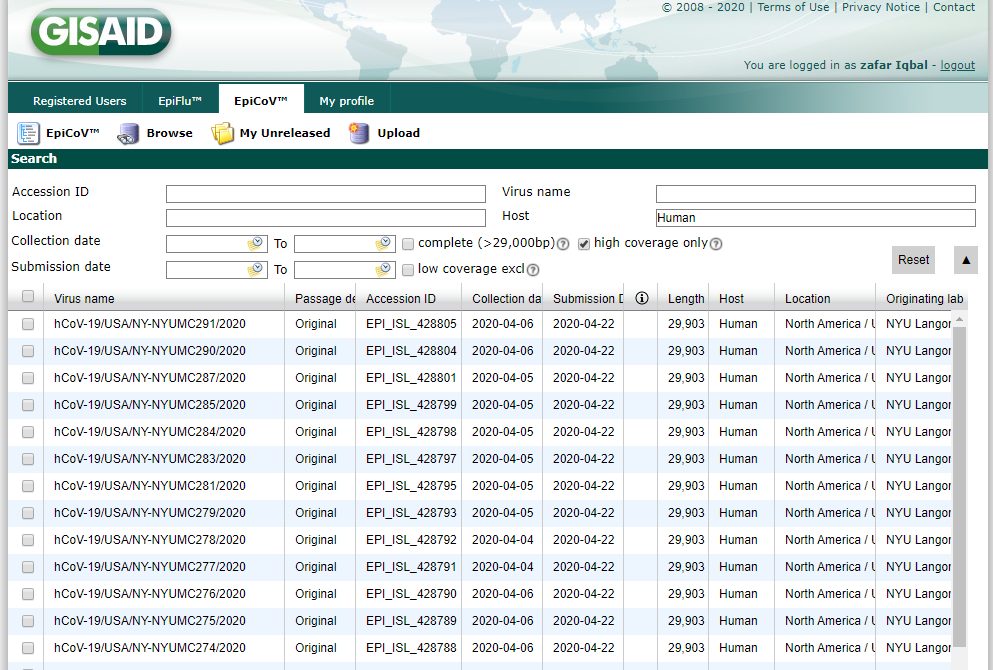


Figure 2: HCoV database

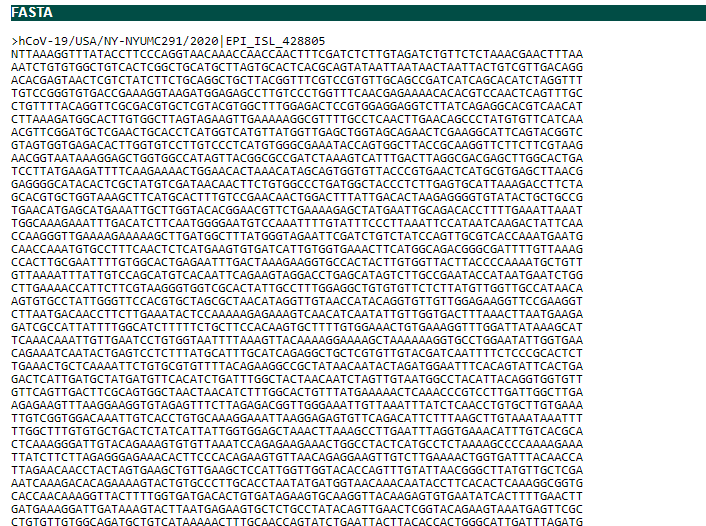


Figure 3: Genome sequence sample

**Toolkit Used to Process Genome Sequences:**

Nextstrain is an open-source project which aims at providing efficient tools for the analysis of pathogen genome data. Their goal is to aid in epidemiological understanding so that appropriate strategy could be adopted to counter the outbreak of diseases caused by viruses. One such toolkit is “Augur” which is a Unix based modular bioinformatics tool. The sequenced data undergoes different steps like filtering and alignment of the sequences, constructions of the phylogeny, getting a time-resolved tree etc. Nextstain also provides a visualization tool called “Auspice”. Figure 3 shows how the sequenced data is processed to get a json file which contains a hierarchal information of source and destination nodes which could be used to get a transmission network of the spread of the COVID-19 disease.

A close up of text on a white background

Description automatically generated

Figure 4 preprocessig of sequcenced data

**Methodology**

We collected the COVID-19 transmission node information in two ways: both manual and automated. For these purposes, we developed and used two Web Crawlers using Java and Python languages. We ran these Web Crawlers on local and remote servers. For the manual part we ran the first Web Crawler on a local Linux server using some specific search keywords such as: COVID-19, coronavirus, and transmission available in the paper or news titles. We initially selected about 375 journal papers, features and news available online. Then we manually reviewed these sources and gathers hundred plus COVID-19 transmission nodes along with source, destination and time information. We also retrieved about hundred stories pertain to these transmission nodes. However, due to page limitation of this report we are able to present here only a few of those that cover from December 2019 to February 2020.

For the automated part, we ran our Web Crawler on cloud-based Linux server in order to extract the COVID-19 transmission networks displayed on the gisaid.org website. From there we imported a .json file with more than 800 active node information from where we extracted their source and destination locations and infection time information of the COVID-19 genome sequences.

Finally, we build two transmission networks: one using the manually extracted transmission nodes that we call our benchmark graph; and another using the automated extracted COVID-19 genome sequences that we call our secondary graph. Finally, we compare and contrast the secondary graph with our benchmark graph. Using these graphs, we also drew some weekly transmission reports, presented by drawing, snapshots, animated presentation to display the transmission sequences of the COVID-19 transmission networks.

**Transmission Graph**

Figure 4 shows a dummy transmission directed graph. In our case, a node refers to a country name and an edge refers to the virus’s transmission from one country to another in one specific date.

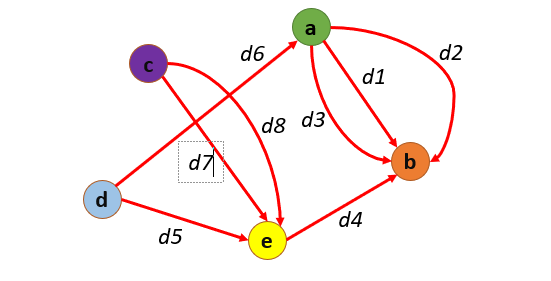


Figure 5: Transmission graph

**Problem Definition**:

Given:

(i) Two graphs with the same nodes and different edge (transmission) sets

(ii) node correspondence

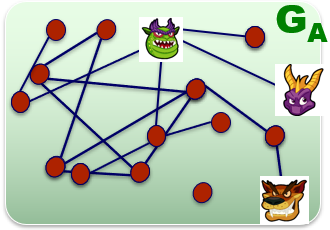
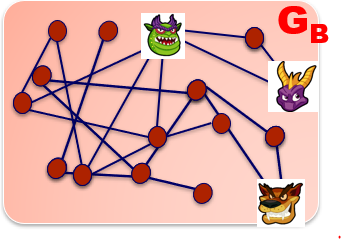
Find:

Edge-based similarity score

s [0,1]

where s = 0: GA < > GB

s = 1: GA = = GB

**Problem Formulation**

*Given:*

Where , for each

Variables:

* from (3-D adjacency matrix [location, location, date])
* (3-D adjacency matrix [location, location, date])

Objective:

Constraints:



**Algorithms**

**Implementation Details**

**Validation Metrices**

**Findings and Discussion**

**References**

1. Submitting Samples to the WHO Collaborating Centre", WHO Collaborating Centres for Reference and Research on Influenza, Melbourne. Retrieved on 2016-11-08.
2. Genomic epidemiology of hCoV-19. Retrieved May 5, 2020 from <https://www.gisaid.org/epiflu-applications/next-hcov-19-app/>
3. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. [published on January 29, 2020]. N Engl J Med. 2020. <https://www.nejm.org/doi/10.1056/NEJMoa2001316>
4. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020. <https://doi.org/10.1038/s41586-020-2012-7>.
5. Qifang Bi, Yongsheng Wu, Shujiang Mei, et al. Epidemiology and transmission of COVID-19 in Shenzhen China: analysis of 391 cases and 1286 of their close contacts. medRxiv. 2020; (published online March 4.) (preprint). DOI: 10.1101/2020.03.03.20028423
6. Hussin A R, Siddapa N. B. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Retrieved on March 13, 2020 from <https://www.sciencedirect.com/science/article/pii/S0896841120300469>
7. Hu, Z., Song, C., Xu, C. *et al.* Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci. China Life Sci.* **63,**706–711 (2020). <https://doi.org/10.1007/s11427-020-1661-4>
8. Lai C C, Shih T P, Ko, W C, Tang, H J, and Hsueh P R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *International Journal of Antimicrobial Agents*, 55(3),March 2020.
9. Ying Liu, Albert A Gayle, Annelies Wilder-Smith, Joacim Rocklöv, The reproductive number of COVID-19 is higher compared to SARS coronavirus, Journal of Travel Medicine, Volume 27, Issue 2, March 2020, taaa021, <https://doi.org/10.1093/jtm/taaa021>
10. Yan Bai, Lingsheng Yao, Tao Wei, et al. Presumed asymptomatic carrier transmission of COVID-19. Retrieved from <https://jamanetwork.com/journals/jama/fullarticle/2762028>
11. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Last Retrieved on April 30, 2020 from https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

# COVID-19 CORONAVIRUS PANDEMIC. Last Retrieved on April 30, 2020 from https://www.worldometers.info/coronavirus/#countries

# Coronavirus Disease 2019 (COVID-19). Last Retrieved on April 30, 2020 from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html

# COVID-19 pandemic in mainland China. Last Retrieved on March 10, 2020 from https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_mainland\_China#cite\_note-AllRegions-10

1. Coronavirus declared global health emergency by WHO. Last Retrieved on March 10, 2020 from https://www.bbc.com/news/world-51318246

# Timeline: How coronavirus got started. Last Retrieved on April 30, 2020 from https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165

1. Aljazeera.com. Timeline: How the new coronavirus spread. Last Retrieved on March 10, 2020 from https://www.aljazeera.com/news/2020/01/timeline-china-coronavirus-spread-200126061554884.html

# COVID-19 pandemic by country and territory: Timeline of first confirmed cases by country or territory. Last Retrieved on April 30, 2020 from <https://en.wikipedia.org/wiki/> 2019%E2%80%9320\_coronavirus\_pandemic\_by\_country\_and\_territory

# COVID-19 pandemic in Thailand. Retrieved on March 23, 2020 from https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_Thailand

1. COVID-19 pandemic in Japan. Retrieved on March 23, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Japan>

# First Case of 2019 Novel Coronavirus in the United States. Retrieved on March 23, 2020 from <https://www.nejm.org/doi/full/10.1056/NEJMoa2001191>

# COVID-19 pandemic in the United States. Retrieved on March 23, 2020 from

<https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_States>

# COVID-19 pandemic in South Korea. Retrieved on March 23, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_South_Korea>

# COVID-19 pandemic in France. Retrieved on March 23, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_France>

# COVID-19 pandemic in Canada. Retrieved on March 24, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Canada>

# COVID-19 pandemic in Malaysia. Retrieved on March 24, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Malaysia>

# COVID-19 pandemic in Australia. Retrieved on March 24, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Australia>

# COVID-19 pandemic in California. Retrieved on March 24, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California>

# COVID-19 pandemic in Germany. Retrieved on March 24, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Germany>

# COVID-19 pandemic in the United Arab Emirates. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_Arab_Emirates>

# COVID-19 pandemic in India. Retrieved on March 25, 2020 from

<https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India>

# COVID-19 pandemic in Italy. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Italy>

1. First Italian dies of coronavirus as outbreak flares in north. Retrieved on March 25, 2020 from <https://www.reuters.com/article/us-china-health-italy-idUSKBN20F0UI>
2. COVID-19 pandemic in Illinois. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Illinois>
3. COVID-19 pandemic in the United Kingdom. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_Kingdom>
4. COVID-19 pandemic in Russia. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Russia>
5. COVID-19 pandemic in Spain. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Spain>
6. COVID-19 pandemic in Sweden. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Sweden>
7. COVID-19 pandemic in Florida. Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Florida>
8. COVID-19 pandemic in Georgia (U.S. state). Retrieved on March 25, 2020 from <https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Georgia_(U.S._state)>