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

CSc 8550 Advanced Algorithms – Final Project Report

Javad Rafiei Asl, Mokter Hossain, 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 [14, 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)

Advanced Algorithms Project Report

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 than its predecessor owing to its high potential of community transmission.

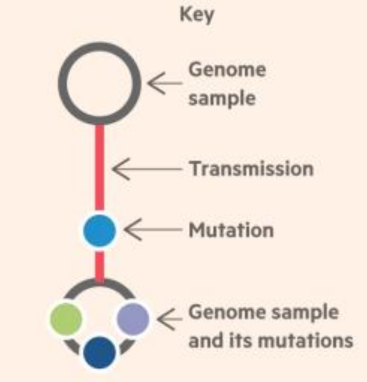


Figure 1. virus mutation

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].

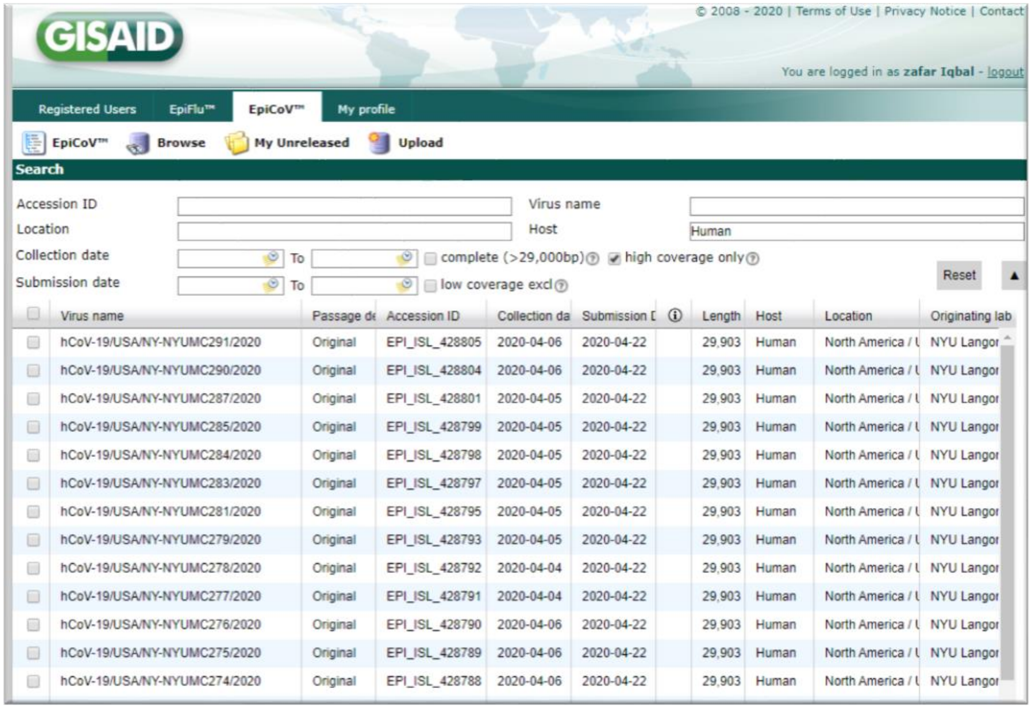
On 31 January, two members of a family of Chinese nationals staying in a hotel in York, United Kingdom, one of whom studied at the University of York, became the first confirmed cases of COVID-19 in the UK [35]. On 31 January 2020, the first cases of COVID-19 were also confirmed in Russia, Spain, and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

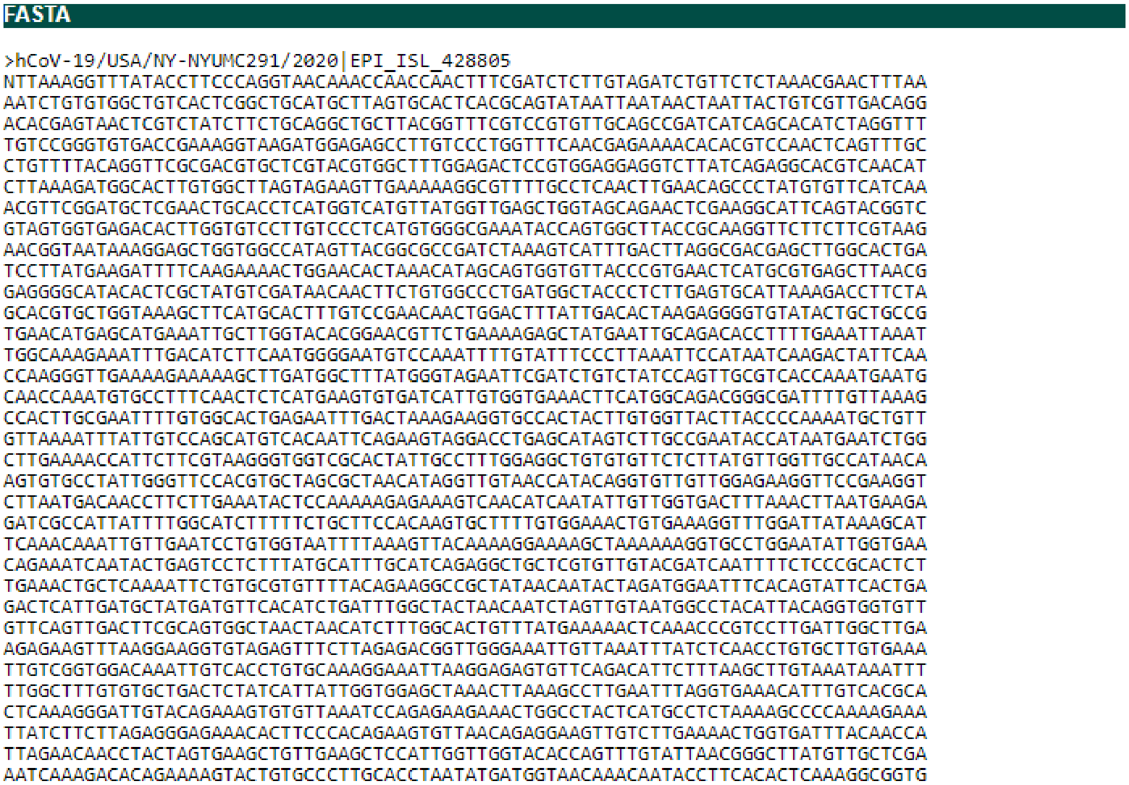
One such repository of Genome sequences is GISAID, which is a public-private venture ship between the government of Germany and non-profit organizations. This initiative promotes the rapid sharing of data about the influenza viruses such as SARS-Cov-2. The data is freely available for researchers all around the globe. Figure 2. shows a screenshot of genome sequences database of SARS-Cov-2 available at GISAID repository. In figure 3., 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.


GISAID
© 2008 - 2020 | Terms of Use | Privacy Notice | Contact
You are logged in as zafar Iqbal - logout
Registered Users
EpiFlu™
EpiCoV™
My profile
EpiCoV™
Browse
My Unreleased
Upload
Search
Reset
▲

Accession ID
Virus name
Location
Host: Human
Collection date To ☐ complete (>29,000bp) ☑ high coverage only
Submission date To ☐ low coverage excl

☐	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating lab
☐	hCoV-19/USA/NY-NYUMC291/2020	Original	EPI_ISL_428805	2020-04-06	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC290/2020	Original	EPI_ISL_428804	2020-04-06	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC287/2020	Original	EPI_ISL_428801	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC285/2020	Original	EPI_ISL_428799	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC284/2020	Original	EPI_ISL_428798	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC283/2020	Original	EPI_ISL_428797	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC281/2020	Original	EPI_ISL_428795	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC279/2020	Original	EPI_ISL_428793	2020-04-05	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC278/2020	Original	EPI_ISL_428792	2020-04-04	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC277/2020	Original	EPI_ISL_428791	2020-04-04	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC276/2020	Original	EPI_ISL_428790	2020-04-06	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC275/2020	Original	EPI_ISL_428789	2020-04-06	2020-04-22		29,903	Human	North America / U	NYU Langor
☐	hCoV-19/USA/NY-NYUMC274/2020	Original	EPI_ISL_428788	2020-04-06	2020-04-22		29,903	Human	North America / U	NYU Langor

Figure 2. HCoV database



```
>hCoV-19/USA/NY-NYUMC291/2020|EPI_ISL_428805
NTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
AATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
CTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACAT
CTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAG
AACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACG
GAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTA
GCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAAT
TGGCAAAGAAATTTGACATCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTGTTAAAG
CCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTT
GTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACA
AGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGT
CTTAATGACAACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAAAGGTTTGGATTATAAAGCAT
TCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAA
CAGAAATCAATACTGAGTCCTCTTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCT
TGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCCGCTATAACAATACTAGATGGAATTTCACAGTATTCACTGA
GACTCATTGATGCTATGATGTTCACATCTGATTTGGCTACTAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTT
GTTCAGTTGACTTCGCAGTGGCTAACTAACATCTTTGGCACTGTTTATGAAAAACTCAAACCCGTCCTTGATTGGCTTGA
AGAGAAGTTTAAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTATCTCAACCTGTGCTTGTGAAA
TTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGTTCAGACATTCTTTAAGCTTGTAAATAAATTT
TTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTTAGGTGAAACATTTGTCACGCA
CTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCCTACTCATGCCTCTAAAAGCCCCAAAAGAAA
TTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTTAACAGAGGAAGTTGTCTTGAAAACTGGTGATTTACAACCA
TTAGAACAACCTACTAGTGAAGCTGTTGAAGCTCCATTGGTTGGTACACCAGTTTGTATTAACGGGCTTATGTTGCTCGA
AATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATACCTTCACACTCAAAGGCGGTG
```

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. Nextstrain also provides a visualization tool called “Auspice”. Figure 4. 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.

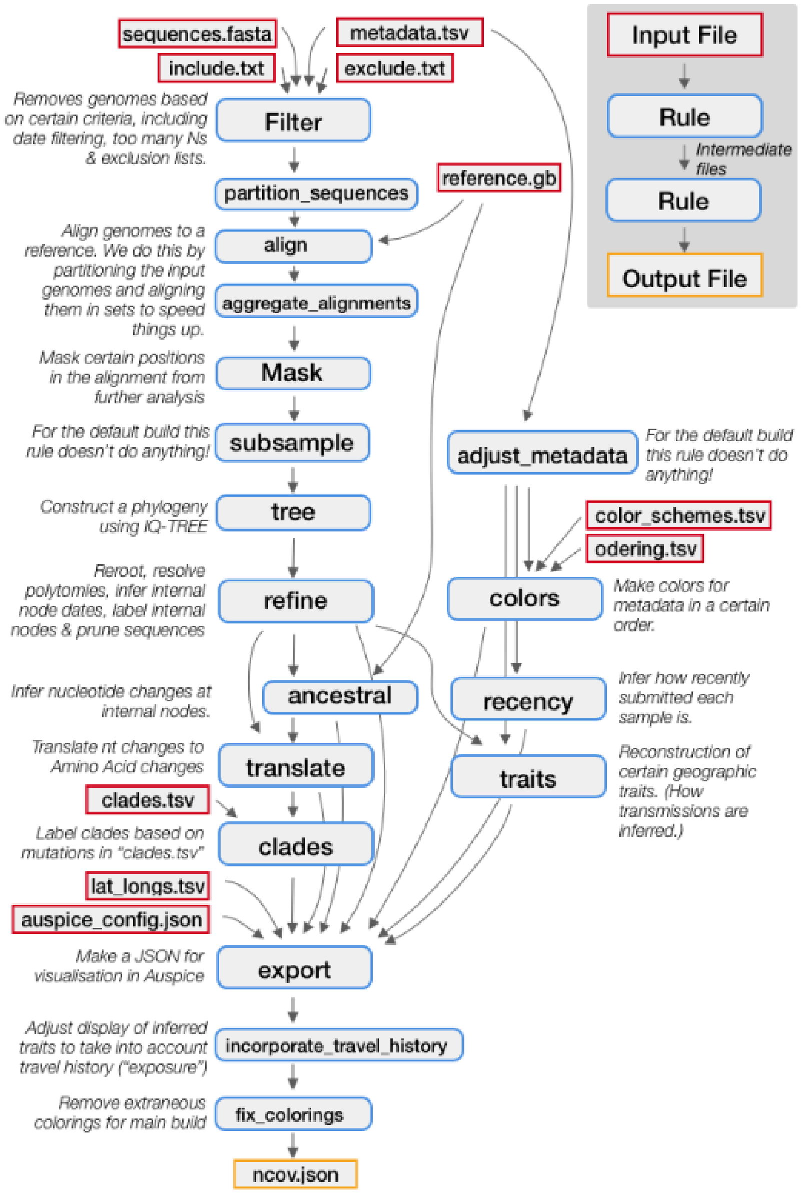

sequences.fasta
metadata.tsv
include.txt
exclude.txt
Input File
Rule
Intermediate files
Rule
Output File
Removes genomes based on certain criteria, including date filtering, too many Ns & exclusion lists.
Filter
partition_sequences
reference.gb
Align genomes to a reference. We do this by partitioning the input genomes and aligning them in sets to speed things up.
align
aggregate_alignments
Mask certain positions in the alignment from further analysis
Mask
For the default build this rule doesn't do anything!
subsample
adjust_metadata
For the default build this rule doesn't do anything!
Construct a phylogeny using IQ-TREE
tree
color_schemes.tsv
odering.tsv
Reroot, resolve polytomies, infer internal node dates, label internal nodes & prune sequences
refine
colors
Make colors for metadata in a certain order.
Infer nucleotide changes at internal nodes.
ancestral
recency
Infer how recently submitted each sample is.
Translate nt changes to Amino Acid changes
translate
traits
Reconstruction of certain geographic traits. (How transmissions are inferred.)
clades.tsv
Label clades based on mutations in "clades.tsv"
clades
lat_longs.tsv
auspice_config.json
Make a JSON for visualisation in Auspice
export
Adjust display of inferred traits to take into account travel history ("exposure")
incorporate_travel_history
Remove extraneous colorings for main build
fix_colorings
ncov.json

Figure 4. preprocessing of sequenced 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 5. shows a dummy graph of transmission network. In this 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. Also, it is possible to have several transmissions from the same source to the same destination in different time slots as show shown in figure 5. In this project, we have generated two transmission graphs: An actual transmission graph from real stories, a predicted transmission graph from GISAID website. We aim to verify the predicted graph based on the real transmissions in the actual graph.

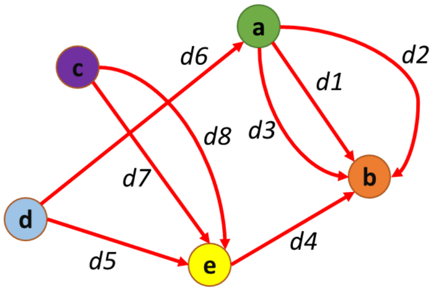


Figure 5. Transmission graph

Problem Definition:

In terms of the problem definition, we propose the "known node correspondence" problem as defined in the following statement and shown in Figure 6.

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: $G_A <> G_B$ && s = 1: $G_A == G_B$

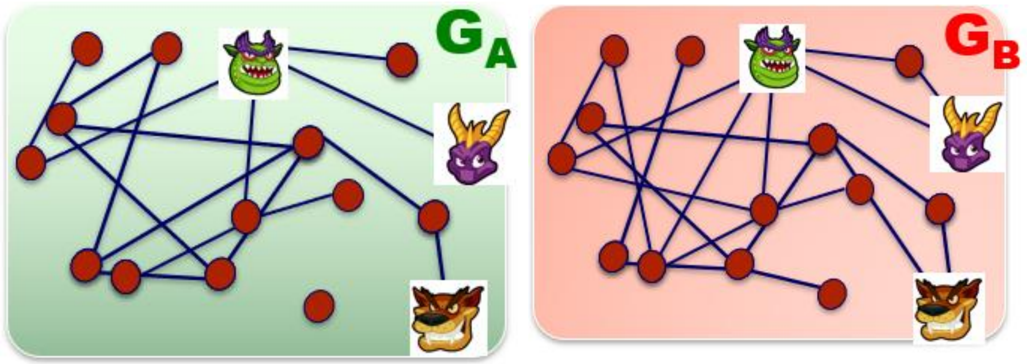


Figure 6. Known Node Correspondence Problem

The Proposed Algorithms

In this section, we have proposed two different algorithms in order to verify transmissions of the predicted (GISAID) graph using the actual transmission graph obtained by different real stories. The first algorithm has employed the adjacency matrix idea in its calculation, thus the first part of this section has concentrated on the adjacency matrix. In the following, both algorithms are investigated in detail.

Adjacency Matrix

An adjacency matrix is a square matrix extracted from a finite graph to represent the edges (transmissions) between the nodes (countries) of the graph. Considering $A_G = [a_{ij}]$ as the adjacency matrix of the transmission graph G, each element of this matrix indicates a virus transferring from a source country to a destination country in a specific time interval. The value of each element of the adjacency matrix can be in range $[0 , +\infty]$ where 0 implies to no transmission between two countries and values more than zero refers to number of transmissions between two countries in a specified time period.

Adjacency Matrix Difference Algorithm

The main idea behind this algorithm is that each transmission from the same source to the same destination appears inside a specified slot of time for both actual and predicted graphs. Therefore, the algorithm divides the timeline of both graphs to distinct time slots with the same length of δ as shown in figure 7. Then, a subgraph with the same time slot is captured from both actual and predicted transmission graphs. In the next step, the algorithm generates the adjacency matrix for each subgraph and then calculates a minor accuracy based on the normalized difference of the actual and predicted adjacencies. Ultimately, the final accuracy is computed by averaging minor accuracies considering different time slots. Figure 8. depicts the general procedure of the algorithm in pseudo code format.

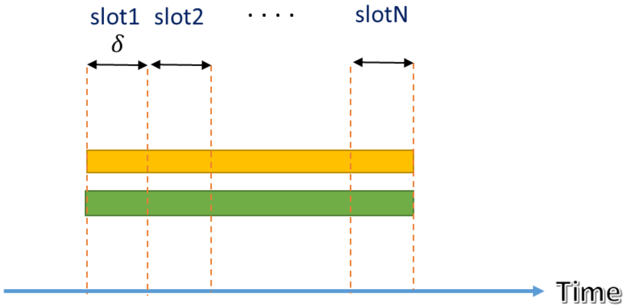
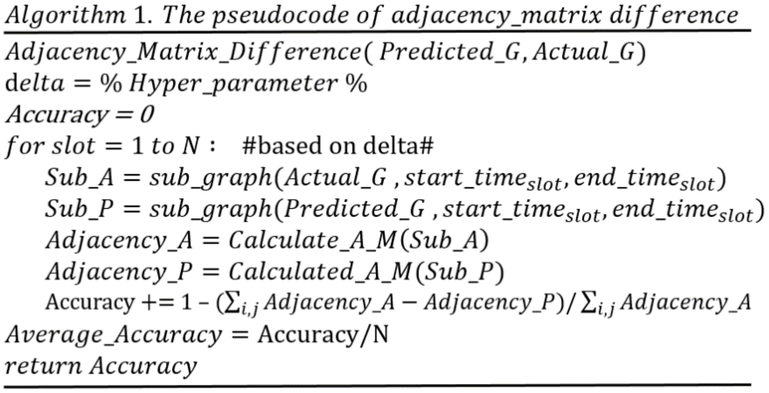


Figure 7. Partitioning of timelines to different slots



```
$Algorithm\ 1.\ The\ pseudocode\ of\ adjacency\_matrix\ difference$
$Adjacency\_Matrix\_Difference(\ Predicted\_G, Actual\_G)$
delta = % $Hyper\_parameter$ %
$Accuracy = 0$
$for\ slot = 1\ to\ N$ :    #based on delta#
    $Sub\_A = sub\_graph(Actual\_G\ , start\_time_{slot}, end\_time_{slot})$
    $Sub\_P = sub\_graph(Predicted\_G\ , start\_time_{slot}, end\_time_{slot})$
    $Adjacency\_A = Calculate\_A\_M(Sub\_A)$
    $Adjacency\_P = Calculated\_A\_M(Sub\_P)$
    Accuracy += 1 – $(\sum_{i,j} Adjacency\_A - Adjacency\_P) / \sum_{i,j} Adjacency\_A$
$Average\_Accuracy$ = Accuracy/N
$return\ Accuracy$
```

Figure 8. The pseudo code of adjacency matrix

Range Based Coverage Algorithm

Instead of using the time slot, the second algorithm utilizes the time range idea covering a specific actual transmission considering epsilon (ε) days before and after the actual date. In other words, for each actual transmission, the algorithm looks for the transmission in the predicted graph within a range with $2 * \varepsilon$ as the length and the transmission date as the center of the range as shown in figure 9. The algorithm aims to calculate the confusion matrix based on the availability of the actual/predicted transmissions in the predicted/actual transmission graph. Considering the figure 10. as a typical confusion matrix, true positive (TP) factor indicates the availability of an actual transmission in the predicted graph, false negative (FN) factor refers to the lack of such a transmission. On the other hand, false positive (FP) factor implies to the lack of availability of a predicted transmission in the actual graph and true negative (TN) indicates to the lack of availability of other transmissions in both actual/predicted graphs.

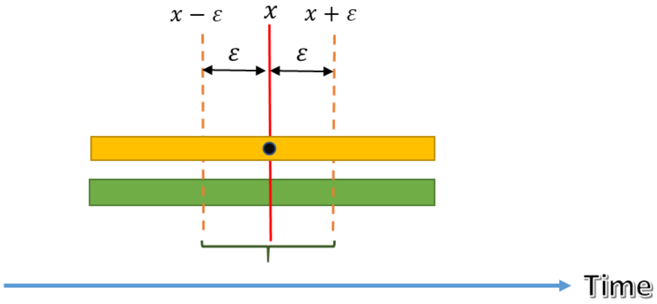
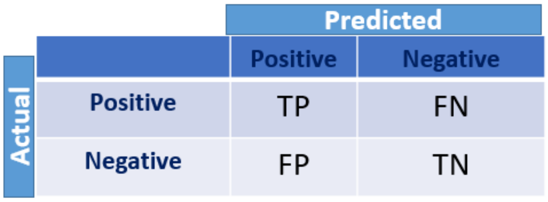


Figure 9. Selecting a covering range for actual



		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Figure 10. Confusion matrix

Experiments and Evaluation

In this section, we aim to depict our experiments and also evaluate them for both devised algorithms in this project. The adjacency matrix difference algorithm has divided timelines of both actual/predicted graphs based on the δ hyper-parameter and then calculate minor accuracy for each time slot in which final accuracy was achieved by averaging minor accuracies. The figure 11. shows the average accuracy based on different values of δ hyper-parameter. For bigger δs, the average accuracy achieves higher performance as the length of time slots are increasing, however the predictions would miss their validities for large δs. Therefore, we have considered $\delta = 3$ as a logical number of days for each time slot where the average accuracy = 25%.

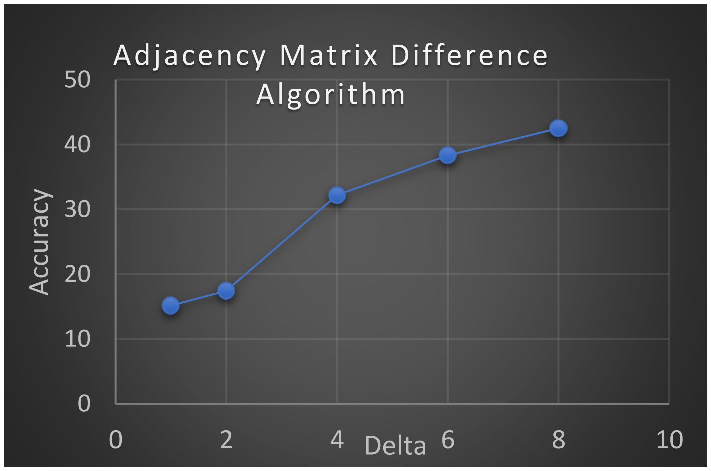


Figure 11. The adjacency matrix difference algorithm performance

The range based coverage algorithm employs a time range covering the input transmission based on ε hyper-parameter as shown in figure 9. The algorithm aims to calculate the confusion matrix in which two important metrics of sensitivity and specificity can be calculated using equation (1) formulas. The figure 12. depicts performance of the predicted transmission graph using these two metrics based on ε hyper-parameter. For larger εs, both of metrics obtains better performance as length of the range is increasing, however as explained already, the predictions would miss their credits for such εs. Therefore, $\varepsilon = 2$ can be used as a rational length for the range of this algorithm. Considering this value, the sensitivity of the algorithm is equal to 28% that is slightly better than the average accuracy of the first algorithm. On the other hands, the high performance for specificity metric (~ 90%) shows that the GISAID graph doesn't predict unavailable transmissions in the actual transmission graph.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad \&\& \quad \text{specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (1)$$

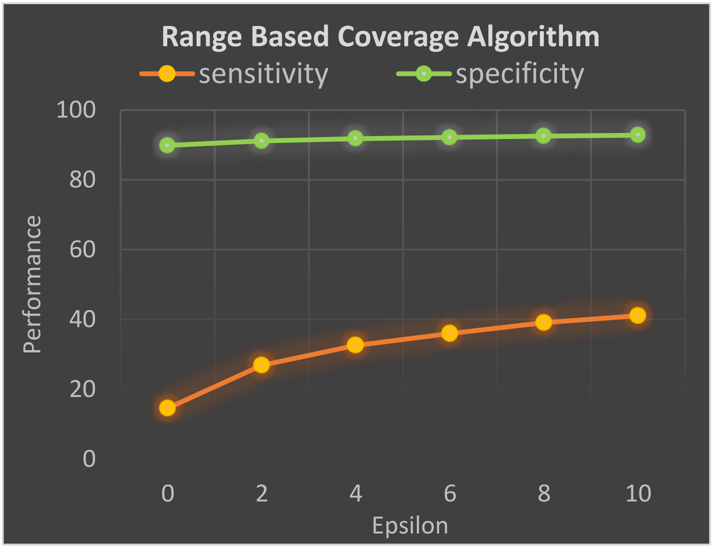


Figure 12. The range based coverage algorithm performance

Discussion and Future work

In this project, we have perused three major facts: 1.) Gathering manually sequences of transmission and stories behind them from trustful resources, 2.) Computing predicted transmission network based on genome knowledge via GISAID website, and 3.) Designing two different algorithms to verify how COVID-19 outbreak occurred. As future work, the following approaches are suggested by our team:

- Enriching actual transmission graph by crawling trustful resources using NLP techniques.
- Designing more complicated algorithms such as DeltaCon [41] and CutDistance [42].
- Comparing the designed algorithms with other state-of-the-art methods.
- Exploring other epidemical repositories to investigate different types of predictions.

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

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

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

14. 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

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

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

17. 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

18. 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

19. COVID-19 pandemic in Thailand. Retrieved on March 23, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Thailand

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

21. 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

22. 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

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

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

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

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

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

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

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

30. 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

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

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

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

33. 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

34. COVID-19 pandemic in Illinois. Retrieved on March 25, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Illinois
35. 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
36. COVID-19 pandemic in Russia. Retrieved on March 25, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Russia
37. COVID-19 pandemic in Spain. Retrieved on March 25, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Spain
38. COVID-19 pandemic in Sweden. Retrieved on March 25, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Sweden
39. COVID-19 pandemic in Florida. Retrieved on March 25, 2020 from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Florida
40. 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)
41. Koutra, Danai, Joshua T. Vogelstein, and Christos Faloutsos. "Deltacon: A principled massive-graph similarity function." In Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 162-170. Society for Industrial and Applied Mathematics, 2013.
42. Klopp, Olga, and Nicolas Verzelen. "Optimal graphon estimation in cut distance." Probability Theory and Related Fields 174, no. 3-4 (2019): 1033-1090.