COVID-19 - A Graph Network Approach

Nibhrat Lohia  
_Southern Methodist University_, nlohia@smu.edu

Rajesh Satluri  
_Southern Methodist University_, rsatluri@smu.edu

Suchismita Moharana  
_Southern Methodist University_, smoharana@smu.edu

Venkat Kasarla  
_Southern Methodist University_, vkasarla@smu.edu

Follow this and additional works at: [https://scholar.smu.edu/datasciencereview](https://scholar.smu.edu/datasciencereview)

Part of the Data Science Commons

**Recommended Citation**

Lohia, Nibhrat; Satluri, Rajesh; Moharana, Suchismita; and Kasarla, Venkat (2021) "COVID-19 - A Graph Network Approach," _SMU Data Science Review_: Vol. 5 : No. 3 , Article 1.  
Available at: [https://scholar.smu.edu/datasciencereview/vol5/iss3/1](https://scholar.smu.edu/datasciencereview/vol5/iss3/1)

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit [http://digitalrepository.smu.edu](http://digitalrepository.smu.edu).
COVID-19 - A Graph Network Approach

Rajesh Satluri\textsuperscript{1}, Suchismita Moharana\textsuperscript{1}, Venkat Kasarla\textsuperscript{1}, Nibhrat Lohia\textsuperscript{2}
\textsuperscript{1} Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
\textsuperscript{2} Adjunct Lecturer, Southern Methodist University, Dallas, TX 75275 USA
\{rsatluri, smoharana, vkasarla\}@smu.edu
nlohia@smu.edu

Abstract. The effects of COVID-19 and its spreads are attributed to various factors. This study uses CDC open-source data on COVID-19 effected population with features ranging from location to ethnicity, to create a Knowledge Graph to measure the similarity between COVID-19 cases and estimate the risk for people likely affected by COVID-19. This data could be used to find correlations between distinct factors, like ethnicity and pre-existing health conditions, to find the vulnerability of a given COVID-19 patient. Using the Jaccard similarity coefficient, in the knowledge graph, we are able to identify and explore relationships between COVID-19 cases as well as predict the vulnerability of general population in a vicinity.

1. Introduction

Found in late 2019, the COVID-19 virus devastated a large portion of the population, spreading quickly because of its high transmissibility, eventually yielding a global pandemic. A Coronavirus is a group of viruses that can affect both animals and humans in various ways. Still, this new strain of Coronavirus, specifically the SARS-CoV-2 strain, affects the respiratory system of humans and can be influenced by previous medical conditions like asthma or lung cancer. Due to the relative novelty of this virus, there are many unknowns within the realm of why one is infected and how to find who is at risk. While 80% of people with COVID-19 recover without special treatment, millions have already died from it. Unlike some diseases, COVID-19 affects all economic backgrounds, albeit some more than others. While lower economic classes have fewer facilities for protection, COVID-19 has found its way into all economic classes because of how easily it can spread and infect people. Higher economic classes can be seen in populated urban centers, which can also influence the spread of COVID-19. This can also be seen in densely populated countries, like India, which recently experienced a massive surge in COVID-19 cases and deaths. This is also due to access to facilities in various places around the world. Many impoverished countries experience issues due to a lack of hygienic protocols and sanitary conditions. A more recent example of a difference in facilities can be seen with the vaccines popping up worldwide. For instance, the most popular vaccines in the U.S. are the Pfizer and BioNTech vaccine, the Moderna vaccine, and the Johnson & Johnson (J&J) vaccine,
all of which have different success rates and side effects [1]. However, the primary vaccines used in other countries like India are the Covaxin vaccine and the Oxford-AstraZeneca vaccine [2]. The Pfizer vaccine in the U.S. has been very successful, with no significant adverse effects and a 95% efficacy in preventing infections [1]. The Covaxin vaccine has an efficacy rate of 78% and can only be administered in 18 and above [2]. Conversely, the Pfizer vaccine can be administered in children of age 12, decreasing the age limit every few months. All these factors influence COVID-19 infection across the world in different ways.

This research needs to know these factors because they can influence the way one is affected by COVID-19. Using the example of location, correlations between population density and increased death rates can be made to find which set of people are more susceptible to the virus. Analyzing these factors is essential because it can help develop these correlations between seemingly unrelated factors, like gender and race, to understand better how one can be affected by the virus. With a good enough analysis of factors, one can develop a system where, given a person with a set of characteristics, their probability of death from COVID-19 can be derived. This information is crucial because it can dictate who must be treated and in order to save as many people as possible.

This research will take several millions of data points from the CDC open-source COVID-19 infection data to build a Graph Database which will eventually feed a Graph Neural Network to build models to find correlations between these factors. In this context, millions of cases are mapped into graph databases by establishing relationships between the factors that influenced the spread. This graph data will be studied in different perspectives through Graph Neural Network models to predict the spread of the virus and the accompanying deaths. Using the same Graph Database, the information can be used to find missing features from the original dataset. In other words, if someone died because of COVID-19 due to a given reason, the Graph Neural Network would predict a trait of the victim, like ethnicity or age. This graph data representation might be useful for fellow researchers to extend the study and thus create a meaningful impact on the fight against COVID-19.

2. Related Work

The outbreak of COVID in late 2019 has been an area of interest for different fields of research. Above all, AI and machine learning play a significant role in predicting and understanding the spread dynamics. This section summarizes past research showcasing how graph theory can be used to make a better decision by describing how the virus spreads, analyzing its severity, and other things.

Wills and Meyer [5] researched graph comparisons which acted as inspiration for our guidelines. Bhapkar et al. (2020) [6] tried to showcase how fast COVID spreads and analyzed if it is one-one, one-many, or one-all in a study done in the early days of the pandemic. Partite graphs and cut sets were used to create a pattern out of past infections and predict when different countries will be affected by the virus by looking into the degree of spread.
In line with the study done by Bhapkar et al., Davahli et al. [7] tried to determine "Expected number of people gets affected by each infectious person" (Rt) which can be used to spread awareness or take preventative measures beforehand. Davahli’s team collected Rt time-series data from January 22, 2020, to November 26, 2020, and experimented with 3 GNNs: Graph-Theory-based Neural Networks (GTNN), Neighborhood-based Neural Networks (NGNN), and the Long Short-Term Memory (LSTM) model to get Rt values for different states. This Rt value is used as an indicator of the severity of the pandemic in different states. Gatta et al. [8] also tried to estimate parameters of the infection spread, such as contact and recovery rate. They took epidemic spread data from Italy and fitted Graph Convolution Neural Network (GCNN) and Long Short-Term Memories (LSTM) to infer the parameters.

While the Bhapkar and Team’s study predicted the start date of COVID in different countries, a study by Mouden et al. [9] focused on clustering more than 200 countries into small buckets based on different factors such as population, death rate, recovery rate, etc. Using the adjacency matrix and its degrees, they used the Laplacian matrix of the input graph to cluster the countries. After clustering, it will be easier to focus on a specific group of countries affected the worst or required the most attention.

COVID-19 patients might be asymptomatic or have similar symptoms to other diseases. To distinguish high-risk COVID-19 patients from similar symptomatic patients, Saha et al. [10] proposed the GraphCovidNet model using a Graph Isomorphism Network (GIN) based model, which is a special category of GNN. They ran image classification models using a graph network to identify potential COVID-19 patients using historic CT scans and chest X-rays. Wu et al. [11] took a similar approach to analyze chest CT scans, but went one step further by estimating the severity of disease spread and identifying lesion areas.

Liang and colleagues [12] also showcased how CT scans can be used for better COVID-19 detection. They did not only consider CT scans but also took time-series CT scan results along with the equipment used for CT scanning information to build a 3D-CNN (Time-series CNN). Liang's research [12], comprised of parts such as feature extraction utilizing 3D-CNN, developing COVID-19 graphs in GCN from features and inputting the graphs to GCN, and transferring learning methods applied in 3D-CNN based on known equipment type.

In measuring the possible impact of COVID variants, sometimes identifying the exact source can help estimate infection level or propagation speed. Li et al. [13] have implemented a Source Identification Graph Convolutional Network (SIGN) aimed to identify the root node. They did extensive experiments on datasets with different infection sizes and got significant results to identify the right source in cases where the infection rate was high. Chen et al. [14] focused on the prevention side by identifying the most influential nodes assuming that high influential persons are most likely to be infected and potentially infect a large number of people. This research collected social network data for people and applied the cluster rank method to identify the most influential nodes.

All the methods mentioned above can increase accuracy if the data is preprocessed. Leung et al. [15] emphasized the importance of clean data. They collected data from heterogeneous sources like age, gender, symptoms, hospitalization status, etc. The researcher looked for any missing values or possible relationships between independent
variables. One important component mentioned in Leung’s study is frequent pattern mining analysis.

Mao et al. (2021) [16] also decided to use multiple data sources to conduct research. They used data collected from the Resident Trajectory Information Database, Confirmed Person Information Database, Information Database of High-risk Groups and Close Contacts, Hospital Fever Information Database, and Mobile Phone Signaling Database then stored it in a graph database format using Neo4j. They found that the graph-based algorithm has high flexibility and can be used effectively by the COVID-19 response team. They also suggested that other countries/regions can use this research in different pandemic stages, be it early, mid, or late stages. The problems were the insufficient intelligence of the data collection method, and the security concern for patient privacy. Mao and Team’s research article gave interest in finding out methods where researchers do not need to worry about the privacy of each confirmed case. So, CDC data for the United States where no patient identification is present is considered in this research paper. In this paper, case data is updated to put a random case ID number to analyze it more to avoid security issues. When data volume is very high, as in the case of COVID related data, Burkhardt [17] emphasizes the scalability issue and details how one can use map reduction to reduce arithmetic operations involved in GNN. He researched a scalable way of identifying significant triangles out of trillion triangles drawn out of data.

To learn more about implementations of Graph Neural Network (GNN) models, Scarselli et al. [18] have made a helpful research paper. They showcased how GNN can be used in website content in Search Engine Optimization (SEO). World Wide Web (WWW) pages are represented in a graph data structure, where pages are graph nodes and links to pages are graph vertices that give a rank to each page based on reliability, importance, and other factors. In addition to PageRank, a node will contain properties like the distinct words used in the page, information on related media and graphic assets, etc. Search engines will utilize the processed graph object and present the audience targeted content. Scarselli and Team’s paper is similar to current research in representing data in a knowledge graph and using other data science approach on it.

Additionally, Su et al. [19] used Bluetooth technology and its signal strength indicator to find if a healthy person was in close contact with an infected person. For proximity detection, they tested the theory using different datasets and different types of phones. Analysis was done for different positions of the phone, like the front shirt pocket, the hand, the back pocket, or the trouser pocket. They compared both the classical estimation theory and the novel hybrid model-based ML approach for proximity detection to understand the complexity and outcome of each method. This Bluetooth technology study shows how ML concepts can be useful in predicting or preventing future breakouts by understanding the underlying pattern or spreading behavior.

This research paper will build knowledge graph and will apply other data science algorithms on it that can predict death based on pre-existing conditions. This paper attempts to solve another aspect of COVID-19 using a graph.
3. Methods

A three-step methodology was used to create the models required to predict the chance of death of a given person or the traits of a person infected by COVID-19. First, data is collected, refined, and has exploratory data analysis performed on it. Second, designing a graph database, where data is loaded by appropriate mapping features to the nodes and establishing relations between the nodes (features). Third, building models using graph Data Science concepts with data gained from the Graph Database.

Data

The data started with 28 million records from the CDC downloaded in the .csv format but is growing continuously. These rows are de-identified patient cases that are available for public use. The dataset has 19 features with a mix of numerical and categorical variables. These are:

1. case_month
2. res_state
3. state_fips_code
4. res_county
5. county_fips_code
6. age_group
7. sex
8. race
9. ethnicity
10. case_positive_specimen_interval
11. case_onset_interval
12. process
13. exposure_yn
14. current_status
15. symptom_status
16. hosp_yn
17. icu_yn
18. death_yn
19. underlying_conditions_yn

Designing a Graph Database

In Relational Database Management Systems (RDBMS), data and relationships between the entities are stored in tables, well suited for transactional data. However, graph databases give lots of advantages over traditional RDBMS databases. Graph Databases yield a lot of flexibility when handling changes with the schema as it does not disrupt current queries. This is very useful for addressing constant changes to the object types, which is not easy to achieve in RDBMS. Nodes, Edges, and properties can be added or dropped to adjust the graph database model without affecting anything else. Graph databases handle large volumes of related data very well. This can be achieved in traditional databases by creating indexes, but there is a lot of overhead in maintaining
the database as the data volume grows. In a graph database, each node maintains its neighboring node's information like pointers in the data structure, which do not need to maintain any global indexing of all the vertices and nodes. This helps the graph database retain its performance even if it grows exponentially, as each node knows its neighbors. When querying the data, graph databases load only related data, which is very advantageous for querying Big Data. The database size grows fast but still retains the real-time response for querying the data.

There are known advantages of using a graph database in many practical problems. Mapping the CDC's COVID-19 infection data into a graph database is very advantageous in drawing relations with different features like county, age-group, ethnicity. It also helps in finding how COVID-19 cases are related to the case identification process, methodology, sex of the victim. And how the victim has been exposed, if the victim was hospitalized. If hospitalized then was the victim in the ICU, what the underlying conditions were, symptoms, and finally if the patient survived. Identifying these case properties and patterns of dependency with other factors is a huge advantage of using a graph database for this research. The data is stored in the Neo4j graph database. The image on the next page (Figure 1) depicts the graph model.

![Figure 1. Graph Design Diagram for CDC Case Data](image)

**Data Science and ML applications**
**Jaccard Similarity Index**

Jaccard Index is a term denoted by $J(A, B)$ and was developed by Paul Jaccard to measure how similar two sets $(A, B)$ are. The index value ranges from 0% to 100%, 100% being the exact match and 0% denoting that the 2 sets are completely different. In other words, Jaccard Similarity is defined ratio of the size of the intersection and the size of the union of two sets. Though the index is easy to interpret, it is very sensitive to small sample sizes as the denominator in this case is very small.

The Jaccard similarity index is calculated as:

\[
J(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

This can be calculated by counting the number of elements present in both the sets $|A \cap B|$, then by counting the total unique numbers present in either of the sets $|A \cup B|$. After that, the $|A \cup B|$ values are divided by $|A \cap B|$ to get the result.

This is mostly used in Natural Language Processing to find the similarity between sentences or documents. This is also used in recommendation system to find the similarity in buying behavior of two individuals based on past purchases. This paper is using the Jaccard Similarity method to find out the similarity between two nodes based on their properties. The similarity index on the nodes of COVID-19 data here is indicating how one node is similar to the other and how it can get affected by looking at the similarity index.

**Graph Neural Networks (GNN)**

The primary purpose of GNN is to learn a suitable representation of graph data called representation learning. Using all the information about the graph, node features, and the connections stored in the Adjacency matrix, GNN outputs new representations called embeddings for each node. These node embeddings contain structural as well as feature information of the other nodes in the graph. This means each node knows something about the other nodes and the connection to the nodes in the context of the graph. The embeddings can finally be used to perform predictions.

Three different predictions can be performed using graph data: node level, edge level, and graph level predictions. For instance, if node-level predictions need to be performed, one can simply use the node embedding of a specific unlabeled node to obtain a prediction.

The picture below (Fig. 2) of GNN shows a graph with four labeled nodes: A, B, C, and D; and one unlabeled node. Then using the embedding vector of the nodes and predict the node labels with it. For a graph-level prediction, one must combine all the node embeddings in a certain way to get a representation of the whole graph. Finally, edge features can also be processed in GNN and will be combined with the node embeddings.
4. Results

Exploratory Data Analysis

To build a model capable of using all the variables, one-hot encoding was performed to convert all categorical variables into binary, integer variables. Preliminary analysis of the dataset reveals there is some multicollinearity between numerous variables. After cleaning the tabular data, Exploratory Data Analysis (EDA) was performed.

In this research, due to the 100 million node limit of the Neo4J database version that was used, a stratified 1 million COVID-19 case sample from 28 million records, which represent the population data, was loaded into the Graph Database by chunking it into 100k records of each .csv file due to the colossal of the original dataset. EDA was then performed on the graph data structure.

The plot below (Fig. 3) indicates that the total number of COVID-19 cases has increased from March 2020 to December 2020. It is also evident that the total number of cases started decreasing from December 2020 to July 2021.

![Fig. 3. Monthly COVID-19 Cases Count (Total)](image-url)
The plot below (Fig. 4) indicated the highest number of cases observed in age groups ranging from 18 to 49 and the second highest cases in 50 to 64.

![Graph showing COVID-19 cases by age group](image)

**Fig. 4.** Monthly COVID-19 Cases Count (For Each Age Group)

The plot below (Fig. 5) indicates that the highest number of cases was observed in CA, NY, and FL.

![Graph showing top 10 COVID-19 death counts by state](image)

**Fig. 5.** COVID-19 Cases Count (State-wise)

After cleaning the tabular data and performing Exploratory Data Analysis (EDA), the data was loaded into the Graph Database. EDA is then performed on the graph data structure. Below (Fig. 6 through Fig. 11) are graphs for each of the relations consisting of sample cases.
Fig. 6. COVID-19 Cases for Alabama

Fig. 7. COVID-19 case relations for Non-Hispanic Ethnicity
Fig. 8. COVID-19 case relations for White Race

Fig. 9. COVID-19 case relations for Jan 2020
Fig. 10. COVID-19 case relations for Abbeville County

Fig. 11. COVID-19 case relations for 0-17 years Age group
This paper has the details of natural orientation for the constructed graph, which means the relationships of the graph were projected the same way that it is stored in the database. Here, the reverse projection did not make much sense to use as it will show the relationships that are pointing towards the case detail node as that node is part of other nodes like county, age group, ethnicity group, or others.

After creating the Knowledge Graph, the Jaccard similarity algorithm in Neo4j was used on the values with a similarity coefficient greater than or equal to 0.7. This cutoff coefficient value was decided to analyze cases which are mostly similar but may not be 100% match, since the cost of false negatives is higher than false positives in this case. The picture below (Fig. 12) depicts the graph representation of one case, 'Case_23777', which belongs to the Washoe County of Nevada State and race 'White'.

![Fig. 12. Jaccard Similarity graph for a sample case](image)

To showcase how the research can be used in future to identify COVID patient’s survival status, cases with death_yn value as ‘Missing’ or ‘Unknown’ was used as test data. And it was done based on its Jaccard Similarity score with other confirmed cases. The figure below (Fig. 13) depicts the Jaccard Similarity for Case_6265, whose death_yn value is ‘Missing’, with respect to other patients whose death_yn status is known. Based on the related nodes (with cutoff 0.7), it can be concluded if Case_6265 will survive. The Jaccard Similarity cutoff can be adjusted based on different use cases to get a better prediction of the patient’s survival chances.
5. Discussion

In this study, the nuances of the spread of COVID-19 were carefully observed. These included the rate at which it spread, who it is more likely to spread to, and how the spread can be predicted. This aspect of COVID-19 is very important to understand because understanding the spread of a virus is fundamental in being able to mitigate its effects. As the virus takes more and more lives, it becomes increasingly important to be able to predict its movements and protect those who need it the most. The Knowledge Graphs and Jaccard similarity used in these studies predict the spread of COVID-19 and its patterns. The idea can be extended to other applications as well.

Because of the nature of the study, the audience is crucial to the viability and usage of our solution. In this study, Knowledge Graphs and Jaccard Similarity was used to identify relationships between various COVID-19 cases, finding patterns in
circumstance, like with county or age group. Such information is crucial in healthcare policy making. For instance, Knowledge Graph can be used to find counties that are at higher risk and bring more medical support to those counties in advance, leading to fewer casualties as a result of COVID-19. On a more individual level, they could use the information to find age groups, ethnicities, or underlying health conditions that makes someone at risk and create a priority order for hospitals so that those patients are the first to get treated, once again, saving as many lives as possible. Health officials can also input the information of those who are not infected by COVID-19 into the Knowledge Graph to find if any person is more vulnerable than another. By factoring in the data of neighboring states or counties, health officials could also find which areas are likely to receive a surge in COVID-19 cases as a result of community spreading.

In this study, a major challenge was being able to load the data available. The CDC library held 28 million records (and counting), which would have been very useful, but because of the constraint on the number of nodes in the database version used, not all this data could be loaded. To overcome this, a stratified sample of 1 million records was used as the representation of the population data. To anyone who wishes to repeat or extend this study, if space and processing power are not an issue, it would be advised to load a full dataset rather than a stratified sample to achieve the most accurate results. Loading this massive amount was another challenge by itself. Memory was a constant issue, so the data had to be chunked into ten chunks of 100,000 records each to overcome the limitation.

After completion of the Knowledge Graph, some interesting patterns within the data emerged. The data shows that the maximum number of cases were reported in the month of December in 2020 from New York, California, and Florida. Most interestingly, the maximum number of cases were from the age group between 18 and 49 years of age, which was unexpected. Most would initially expect that COVID-19 would affect the older age groups like senior citizens over middle-aged people, but the data proved otherwise.

Certain fields of study can use this study and take it to a deeper level. For instance, machine learning or neural network experts could use GNN models on top of this Knowledge Graph to predict missing relationships, nodes, or subgraphs and further analyze the data. Using GNN, one could find not only relationships between the data, but also the possible survival chance of a given COVID-19 patient, and this could prove very useful within the medical field. This study could also be used in the field of epidemiology, the study of the transmission of viruses in a population. Using the methods of Knowledge Graphs and Jaccard similarity, epidemiologists can more easily predict factors that influence the spread of a virus and could further the epidemiological field significantly.

Our data was from the open-source CDC library, which logs information like race, ethnicity, age group, and preexisting health conditions, but purposefully leaves out personal information like specific ages or names. By this logic, the study is ethically sound as it does not infringe on the personal privacy of any patient whose information was taken and used in the database. However, there are some people who are not
comfortable sharing their health conditions, age group, or county location, and this makes ethics more complicated as it involves opinion.

This research is designed such that there is no need for any personal information of the population to identify the community spread or individual vulnerability. Therefore, anyone who would like to use the knowledge graph and the data science methodologies implemented in this research can use anonymous information for their studies or should get consent from their sampled population before starting their study.

6. Conclusion

To conclude, with the use of Knowledge Graphs and data science methods like Jaccard similarity, we were able to effectively identify relationships between COVID-19 cases and their characteristics. The ability for someone to use a Knowledge Graph to effectively find these relationships within the data is a success by itself because building the graph data structure while preserving all the relationships within the data and loading such a massive amount of data is a challenge. The relationships learnt can be used in policy making at various levels. Using a Knowledge Graph for this data is beneficial because it allows for faster queries of the data within the graph and extendibility of the database without disrupting the current functionality.

Acknowledgments

We would like to express our sincere gratitude to Jacquelyn Cheun, PhD. – Capstone Professor for her continued support and guidance.

References

18. Scarselli, Franco & Yong, Sweah & Gori, Marco & Hagenbuchner, Markus & Tsoi, Ah & Maggini, Marco. (2005). Graph Neural Networks for Ranking Web Pages,. 666-672. 10.1109/WI.2005.67. https://doi.org/10.1109/WI.2005.67

Data Sources

- Center for Disease Control and Prevention (CDC) data for COVID-19
- https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4