Visual Understanding of COVID-19 Knowledge Graph for Predictive Analysis

Seung-Hwan Lim, Junghoon Chae, Guojing Cong, Drahomira Herrmannova,
Robert M. Patton, Ramakrishnan Kannan, Thomas E. Potok
Oak Ridge National Laboratory
lims1, chaej, congg, herrmannovad, pattonrm, kannanr, potokte@ornl.gov

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2 /21/$31.00 USG | DOI: 10.1109/BIGDATA52589.2021.9672055

***Abstract*—This study aims to effectively analyze and visualize the concept to concept network derived from the COVID-19 Open Research Dataset (CORD-19) dataset, where we have more than 48,000 concepts with more than 300,000 relationships between concepts. In analyzing networks, we focus on finding relationship patterns between the coronavirus disease 2019 (COVID-19) concepts and other concepts. Given the node and edge datasets, we construct directional graphs and calculate all pair shortest paths based on multiple edge weight schemes. However, statistical metrics are not sufficient to identify specific relationships represented in the network. Therefore, we also propose a visual analytics approach to effectively understand the knowledge graph. Our highly interactive visual analytics allows users to effectively analyze the evolving graphs and (COVID-19) concept nodes and other nodes related to the COVID-19 nodes. We envision that this study will pave the path to develop strategies to provide more accurate and scalable predictive analysis on knowledge graphs related to CORD19 and other biomedical knowledge graphs.**

***Index Terms*—COVID-19, knowledge graph, visualization**

1. Introduction

Graph is a universal language that describes many physical or social phenomena by representing relationships between constituents. In graph-based analysis, a useful tool is knowledge graph, which is a knowledge base that uses a graph-structured data model or topology to integrate data across domains such as natural language processing [1]. An active area of research related to knowledge graph is reasoning from knowledge graph under the circumstance of missing information such as predictive path queries [2] and predicting (or completing) missing/future links [3]. Moreover, knowledge graphs often exhibit dynamism; new nodes can be created over time and new edges also can be created between a pair of nodes over time. Thus, by considering the dynamism, we may derive additional insights from the knowledge graph. Another important challenge is that predictive reasoning on large graphs often requires special treatments in order to obtain the desired accuracy of prediction [4].

Toward the end, the understanding of the characteristics of graphs can be important to develop an analysis methodology for the target graph. Thus, this study provides the statistical characteristics of the graph in the combination of visual investigation of specific interesting cases in the graph, with the example of coronavirus disease 2019 (COVID-19) knowledge graph derived from publications curated in the Covid-19 Open Research Dataset (CORD-19) dataset [5]. Our study follows lineage of knowledge graph-based studies related to this dataset [6]–[8]. The recent open release of publication relevant to the current SARS-CoV-2 pandemic, CORD19, is an exemplar of rapidly growing scientific knowledge. Since the pandemic has become the dominant topic in society, an urgent need is to reason information from such vast stores of knowledge from the relevant scientific literature growing at incredible rates. Given that the volume of information is easily beyond the capacity of any one person, analysts have been strongly motivated to develop automated knowledge-mining methods and extraction tools [9], [10].

For improving the accuracy of predictive analysis in the knowledge graph, we focus on the distance relationship between all nodes. To obtain a distance relationship between all nodes, we calculate all pair shortest paths (APSP) based on three different edge weight schemes. Based on the APSP results, we analyze the distribution of distance from all nodes to the nodes related to COVID-19 and from the COVID-19 nodes to all other nodes. We want to reveal what concepts are closely connected to the COVID-19 concept in the knowledge graph.

In addition, we introduce a visual analytics approach to understand the knowledge graph related to the distance relationships calculated from APSP. We visualize the concept nodes that are in the shortest paths between concept nodes. Also, our visualize provides interactive features to compare multiple graphs effectively, which is intended to effectively visualize the temporal evolution of the knowledge graph. We will present relevant important prior work in the next section, followed by motivational example related to high-performance computing in Section III; detailed description of the construction process of our knowledge graph in Section IV; summary of analysis results in Section V; analysis

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Robinson Pino, program manager, under contract number DE-AC05-00OR22725.

results in Section VI; and conclusions in Section VII.

2. Our approach

Constructing a knowledge graph from the data

We construct our knowledge graph to represent concept to concept relationships, where concepts are nodes and relationships between concepts are edges. We define concepts as any entities or terms appearing in the Unified Medical Language System[1] (UMLS). The UMLS is a collection of biomedical controled vocabularies and ontologies that links these vocabularies into a single hierarchy. It also provides information about known relations between the concepts (entities and terms) included in the hierarchy. We use the definitions of concept and relation used by the UMLS, that is, a *concept* represents a "meaning" and a *relation* represents both hierarchical and associative relations. Alongside UMLS, the National Library of Medicine (NLM) also maintains SemRep[2], an application which extracts UMLS concepts and relations in the form of semantic triples (called predications in SemRep) from natural text [11].

The COVID-19 Open Research Dataset[3] (CORD-19) is a collection of scholarly articles about the novel coronavirus. We use the SemRep processed version of CORD-19 released by the NLM[4]. More specifically, we use the extracted predications, which are provided in the form of entity-relation-entity triples, along with the information about the publications which these concepts were extracted from, to construct our graph. In order to associate these concepts to papers, we counted the number of times each concept appeared in a paper and associated each paper to the relationship between concept c_x and c_y by the Jaccard similarity score:

$$w_{c_x c_y} = log \frac{|C_x \cap C_y|}{|C_x \cup C_y|}$$

On top of these extracted nodes and edges, we construct a directional graph, where nodes are concepts and links represent relationships that connect concepts. Each link has weights based upon three different schemes:

$$w(u,v) = \begin{cases} 1 & \text{unit weight} \\ log(N/n)/log(N) & \text{log weight (or entropy)} \\ 1/n & \text{inverse,} \end{cases}$$

where N is the total number of papers, and n is the number of papers between concepts and unit weight assigns the weight of 1 to all links, regardless of the number of papers between concepts. The overall idea of entropy and inverse weight schemes is to consider the concepts with a larger number of papers as closer concepts to each other than the concepts with a smaller number of papers. The unit weight scheme is to maximize the impact from the structural property of the graph. We do not utilize the types of relationships associated with each paper in the construction of graph, but we use them in the analysis of graph.

Visual Analytics for Knowledge Graph We develop a specialized visualization for the COVID-19 graphs as shown in Figure 1. The node-edge diagram is based on a force-directed graph drawing algorithm. We set the algorithm so that the nodes with a strong relationship attract each other and other nodes repel each other. In this visualization, the light purple color nodes represent COVID-19 concept nodes, and the orange color nodes indicate bridging nodes connecting the COVID-19 nodes. The thickness of the edges shows the number of connections, and the darkness of the blue edges represents the distance between nodes. In other words, thicker edges depict that the source and target nodes between the edges have more connections than thin edges. It means that there are more papers associated with both concepts. Darker blue edges show that the source and target nodes are closer to each other based on the specific edge weight scheme. For example, in Figure 1, the *IMPACT gene* and *Patients* nodes have the largest connections and *Evaluation procedure* and *SARS-COV-2 vaccination* nodes are closest to each other.

Concept Unique Identifier (CUI)	Name
C5203670	COVID-19
C5203676	2019 novel coronavirus
C5203671	Suspected COVID-19
C5203672	SARS-CoV-2 vaccination
C5203674	Antibody to SARS-CoV-2

TABLE 1: COVID-19 related CUIs.

As we mentioned earlier, we construct multiple graphs based on the three different schemes. We construct two graphs for each scheme where the first graph is created by the edges with timestamps before July 1, 2020, and the second one is made by the entire edges. So, the first graph is a sub-graph of the second one. For example, two graphs of the unit weight scheme are shown in Figure 7. Given the graphs, we select concept nodes related to COVID-19. We used the concepts related to COVID-19 in Table 1. We find the shortest path to each other between nodes involved in COVID-19 and then construct COVID-19 graphs using the shortest paths including COVID-19 and other intermediate nodes.

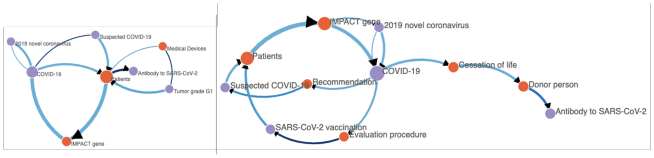


Figure 1: Visualization of COVID-19 Knowledge Graph: Graphs of Inverse Scheme

3. Results

This section describes our analysis on the characteristics of 4 different snapshots of knowledge graphs that we

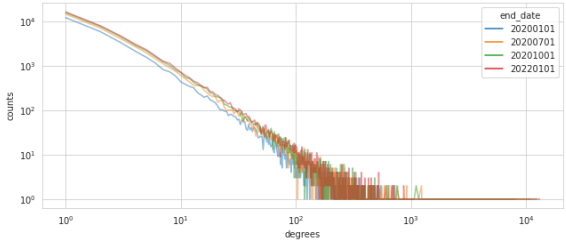
1. https://www.nlm.nih.gov/research/umls/
2. https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html
3. https://www.semanticscholar.org/cord19
4. https://lhncbc.nlm.nih.gov/ii/information/COVID-19.html

cut-off date	number of nodes	number of edges	average density
2020-01-01	34,211	174,555	5.10
2020-07-01	42,400	250,795	5.91
2020-10-01	46,435	293,235	6.31
2022-01-01	48,775	323,715	6.64

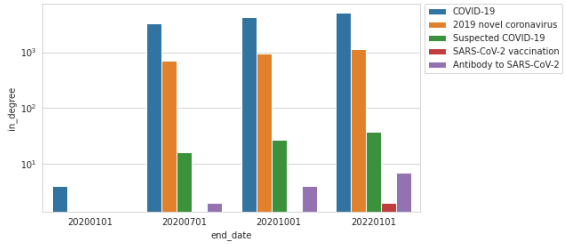
TABLE 2: basic statistics of COVID-19 graph over the time

split based upon cut-off dates. Overall, in our knowledge graph, new concept nodes and new relationships continuously emerge, while a majority of new relationships (or new edges) occur between already connected concept nodes. We confirm that the distance between nodes varies according to weighting schemes, leading to finding different paths, even though the number of hops might be similar. Let us look into detailed analysis results in detail.

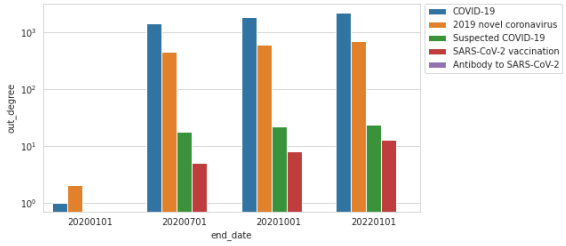
For the visualization results, the graphs using edges before 7/1/2020 have only a few bridging nodes between the COVID-19 nodes, and the concepts are general. On the other hand, the graphs using entire edges have many bridging nodes between the COVID-19 nodes. Also, the concepts are more specific. There are also close distance connections between intermediate nodes.



(a) Degree Distribution of COVID-19 Knowledge Graph



(b) in degree of covid-19 terms



(c) out degree of covid-19 terms

Figure 2: Basic descriptive statics

3.1. Basic descriptive statistics on knowledge graph

Before a high-level understanding of the graph, let us look into basic descriptive statistics on the target graph, such as the size of graph (Table 2, degree distribution of nodes (Figure 2a), degree of popular COVID19 concepts (Figure 2), and top 15 nodes with respect to degree (Figure 3).

Table 2 shows the number of nodes and edges in the constructed COVID-19 knowledge graph according to cut-off dates. Note that the last cut-off date simply intends to include all the nodes and edges in the data. We observe that the number of nodes and edges consistently increase. Also, the average density (# of edges/# of nodes) steadily increases. This trends means that the number of edges are increasing faster than the number of nodes. Thus, we can state that the concepts and relationships between concepts have been consistently increased during the year of 2021 as scientific discoveries progress.

Figure 2a) shows the degree distribution of nodes in our COVID19 knowledge graph. In this analysis, we confirm that the constructed knowledge graph follows power-law distribution since we observe stead decline in log-log plot from all 4 different cut-off dates, though the graph has consistently added new nodes and edges. Thus, we can say that the evolution of knowledge follows similar pattern with other social networks. Also, if we apply analysis techniques, we can exploit this degree distribution pattern.

In the degree distribution, very few nodes have degrees higher than 10^3. Figure 2 shows that the term of COVID-19 is among the high-degree nodes from 2020-07-01. It is natural that COVID-19 related terms have very low degree or they are not exist since this disease is caused by a novel virus discovered in 2019. However, the scientific findings rapidly grow, some of COVID-19 terms such as COVID-19, and 2019 novel coronavirus has become one of the high-degree nodes in the knowledge graph.

Figure 3 shows top 15 high degree terms per cut-off dates. This anslysis shows that “Patient” is the most popular term. Since it is a very generic term, almost all concepts can be connected to this term. However, we observe that more COVID-related terms and terms relevant to similar disease (e.g., severe acute respiratory syndrome) become more popular as time goes. From this analysis, we can see the challenge to analyze COVID-19 knowledge graph related to those high-degree generic terms. For instance, although the amount of new information related to those generic terms will be very low, dropping this nodes will result in dropping a lot of relationships. Thus, we may need a careful treatment to discover useful knowledge related to these high-degree generic terms.

3.2. Semantic distances related to COVID-19 concepts and implications to predictive analysis

This section shows the impact on the number of hops in the shortest path according to weight schemes, which reveals the semantic distances between medical concepts related to COVID-19. As shown in Figure 4 and Figure 5, we observe that the “inverse” weight scheme tends to leading to longer semantic distances between concepts. On the other hand, “log” weight scheme leads to similar semantic distance “unit” weight scheme. We attribute this result to

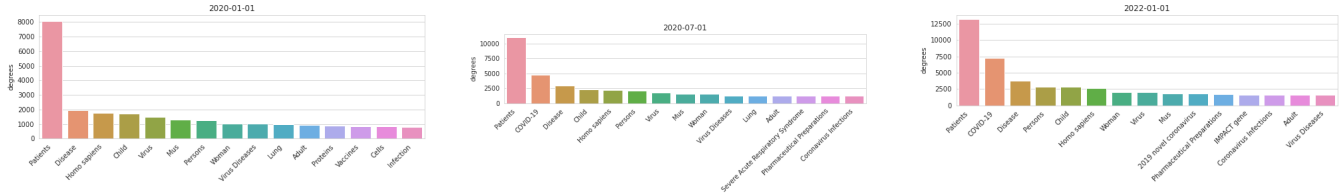


Figure 3: Top 15 degree terms per cut-off dates show that more relevant terms to the disease emerge.

log weight scheme (distance=3.43)				
src name	rel	paper count	weight	dst
Suspected COVID-19	PROCESS OF	344	0.574	Patients
Patients	LOCATION OF	377	0.567	IMPACT gene
IMPACT gene	AFFECTS	3	0.920	Lung
Lung	LOCATION OF	554	0.539	Ultrasonography
Ultrasonography	USES	5	0.883	Sonazoid
unit weight scheme (distance=5)				
src name	rel	paper count	weight	dst
Suspected COVID-19	PROCESS OF	3	1	"Infant, Newborn"
"Infant, Newborn"	LOCATION OF	9	1	IMPACT gene
IMPACT gene	INTERACTS WITH	1	1	Epithelial Cells
Epithelial Cells	LOCATION OF	1	1	Diagnosis
Diagnosis	USES	1	1	Sonazoid

TABLE 3: a path from Suspected COVID-19 to Sonazoid

our treatment in "log" weight scheme, where we normalized edge weights between 0 and 1. However, the actual shortest path can be different between "log" weight scheme based distance and "unit" weight scheme as shown in Table 3. In "inverse" weight scheme, the edge weight can be very small if the number of papers associated with the pair of concepts is large. Thus, an expressive metric that can capture the semantic relationship between concepts will play a critical role, for researchers to find hidden relationships between concepts.

Figure 4 and Figure 5 shows the distribution of distances of incoming (Figure 4) and outgoing paths (Figure 5) of COVID-19 nodes from/to all other nodes, in our constructed COVID-19 knowledge graph. In this analysis, we can find that the distributions of distances of incoming and outgoing paths are not symmetric and the changes of distances reflect the progress of scientific findings. For example, we find that several popular COVID-19 concepts such as 'Suspected COVID-19', and 'SARS-CoV2 vaccination' appear between Jan, 2020 and June, 2020. Thus, they do not have outgoing paths in Jan, 2020 version of COVID-19 concept graph.

We also find that inverse weight scheme shows higher sensitivity to the number of papers and tends to skew toward shorter distances. The entropy-based weight scheme seems to follow more bell-shaped distribution. Unit weight scheme is also useful to find the number of hops between concepts. In this boxplot, we observe that concepts are usually 2 to 3 hops away from COVID-19 concepts. When we look at outliers in the distance from popular COVID-19 concepts, we continue to have concepts far from COVID-19 concepts. It is mainly due to the fact that newly introduced concepts in this COVID-19 concept graph as bio-medical researchers find new concepts or new links with concepts that have not been considered in the context of COVID-19. Also, we can notice that concepts are connected to COVID-19 terms in less than 8 hops in both incoming direction and outgoing direction.

Now, let us discuss the lessons learned from this analysis in order to advance the quality of predictive analysis in semantic distances such as link prediction. An important assumption in link prediction is that semantically or structurally close nodes will have a higher chance to have a future link [12]. To validate the assumption, we analyzed the relationship between semantic distance and structural distance. In order to show the relationship, we analyzed distances between nodes when they have new direct links according to different weighting schemes, as shown in Figure 6. In this analysis, we investigate the number of hops in the shortest paths when we calculate the shortest path based upon three different weighting schemes.

We can interpret that the results on "unit" weighting scheme show the structural changes, which will serve as a baseline for the other two weighting schemes – "log" and "inverse". We find that the log weighting scheme behaves similarly to the unit weight scheme with respect to the previous hops between newly connected nodes - mostly peak around 3 and 4 hops. However, "inverse" weighting scheme tends to connect longer hop nodes. It may imply if we use "inverse" weights as edge attributes, it might be a bit harder to accurately predict future links if we assume already connected nodes can have additional edges. Thus, predicting semantically meaningful new links may need a careful investigation in evaluating the importance of relationships attributes such as edge weights.

In addition, another insight for link prediction from this distance distribution analysis is the challenge of predicting break-through links connects concepts with longer distances, which might be more semantically interesting beyond the accuracy of a link prediction model. This might be a statistically challenging scenario since most of the links are associated with high degree nodes and they will be likely to have more new links. Thus, imbalanced link distribution might be a challenge to predict semantically meaningful or breakthrough links.

3.3. Discovery concepts between COVID-19 nodes

The two graphs in Figure 7 shows the graph using edges before 7/1/2020 (left) and the entire edges (right). As shown in the results, the two graphs are very different. The left graph is much smaller than the right one. The left graph has only three bridging nodes between the COVID-19 nodes and the concepts are general. There are not many specific bridging concepts between COVID-19 before July 1st, 2020. On other hand, the right graph has many bridging nodes between the COVID-19 nodes. Also, the concepts are more

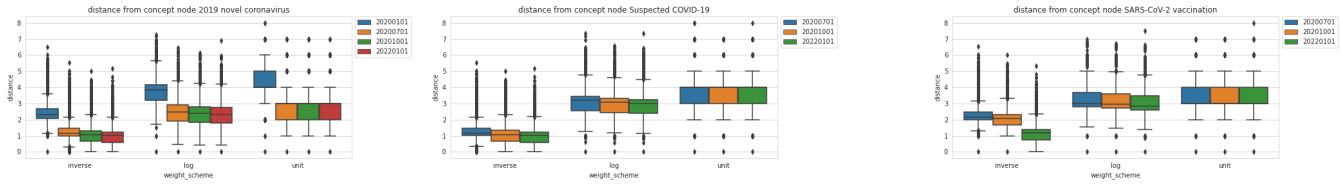


Figure 4: Outgoing distance from COVID19 concept nodes to all other nodes

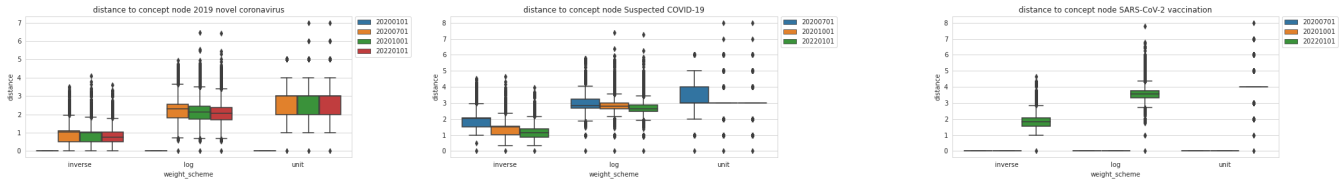


Figure 5: Incoming distance to COVID19 concept nodes from all other nodes

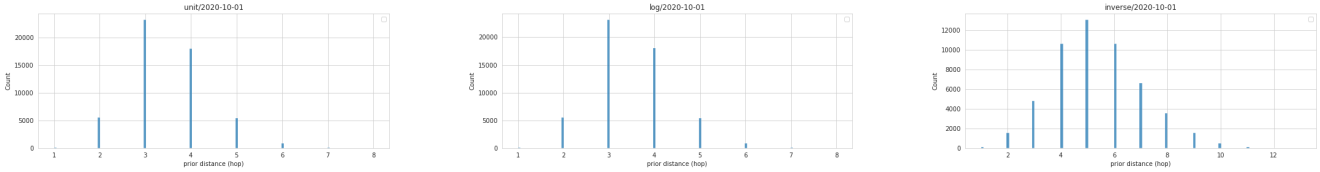


Figure 6: Distances between concepts when they have new links according to weighting schemes

specific than the left graph. There are also close connections between intermediate nodes.

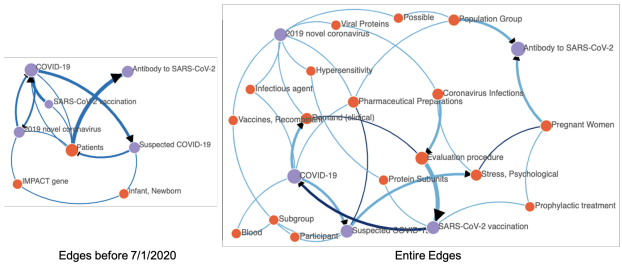


Figure 7: Graphs of unit weight scheme

In this visualization, we also provide interactive functions to compare specific nodes between the two graphs. When users select a specific node of one the two graphs, the node with the same ID of the other graph is selected. Also, the nodes and edges connected to the selected node are highlighted and other nodes and edges are dimmed out. Theses interactive features help users to compare the two graphs efficiently.

We also visualize the COVID-19 graphs constructed by the log scheme. The results are shown in Figure 8. In this case, the two graphs are not different as much as the unit weight scheme case. For the right graph, however, we can see there more connections and the distances between the COVID-19 nodes are shorter than the left graph. This means that more papers related to the multiple COVID-19 concepts are published since July, 2020.

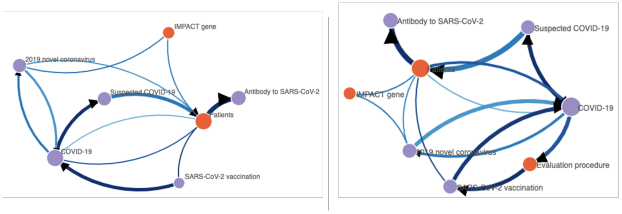


Figure 8: Graphs of log weight scheme

For the inverse scheme, we can find out new specific bridging concept nodes between the COVID-19 nodes in the right graph, such as *Cessation of life* and *Evaluation Procedure*. However, some concepts are still general and the COVID-19 nodes has long distance to each other based on the inverse scheme.

3.4. Discovering nodes closest to the COVID-19 nodes

The distances from the concepts to the COVID-19 nodes are long before July 2020, but the distance decreased a lot when we use the entire edges. It means that many research papers connecting the concepts and COVID-19 concepts published since July 2020. For every single node, we measure the distance to each COVID-19 node in the

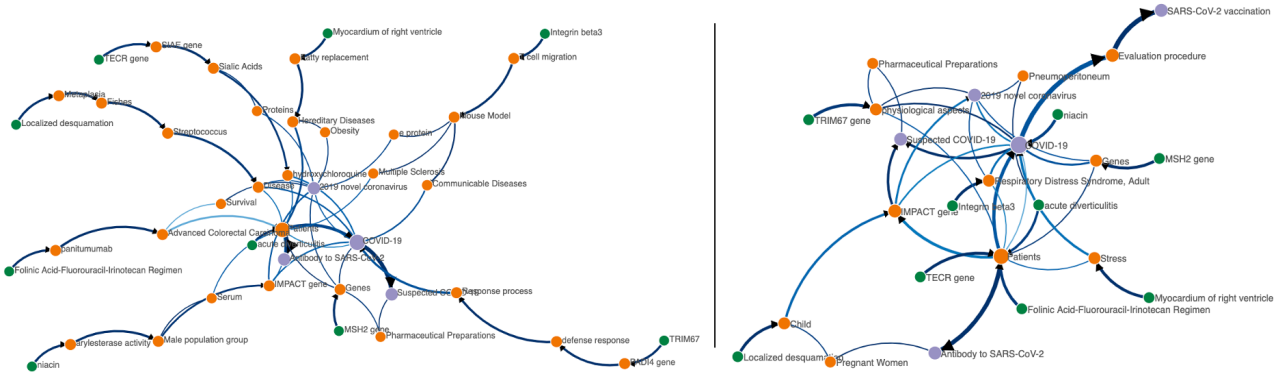


Figure 9: The graph with the edges before July 2020 (left) and the entire edges (right). The green color nodes are closest to the COVID nodes since July. The purple ones are the COVID-19 nodes and the orange ones are the bridging nodes.

graph with the edges before July 2020. Again, we measure the distance in the entire graph in the same manner. For this task, we use the log scheme only. Next, we filter top-k nodes ($k = 10$) with the most reduced length for each COVID-19 node. Then, we select the nodes that appear in common for multiple COVID-19 node cases. Especially, Integrin is a protein that may act as an alternative receptor for SARS-CoV-2 and could be implicated in its transmission and pathology. Also, Niacin is one of the B vitamins. Vitamin potentially improving outcomes and reducing the length of stay in the hospital for COVID-19 patients.

Finally, we construct graphs using three types of nodes. First is the selected nodes, the second is the COVID-19 nodes, and the third is bridging nodes between the first and second type nodes. The results are shown in Figure 9. The left figure shows the graph with the edges before July 2020 and the right one shows the graph with entire edges. The green color nodes are the selected nodes, the purple ones are the COVID-19 nodes, and the orange ones are the bridging nodes. As shown in the results, we can see that green nodes are getting much closer to the COVID-19 nodes. Also, the bridging nodes disappear in the end.

4. Conclusions

This study presented a detailed analysis of an instance of knowledge graph related to publications on COVID-19. Our analysis shows the knowledge graph follows power-distribution in terms of degree of nodes in the graph. The distance between all nodes and identified common COVID19 terms is less than 8-hops. This study also shows the potential of importance of semantically meaningful weighting schemes through visual analysis, which can lead to find semantic distance between concepts and, potentially, more accurate link prediction. Although this study is analysis on a single example, we believe that findings around the example COVID-19 knowledge graph is generic enough to inspire the direction to designing more accurate and meaningful link prediction and to implementing machine learning algorithms on graphs in a scalable manner.

References

[1] W. Jin, M. Qu, X. Jin, and X. Ren, "Recurrent event network: Autoregressive structure inference over temporal knowledge graphs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6669–6683.

[2] W. Xiong, T. Hoang, and W. Y. Wang, "Deeppath: A reinforcement learning method for knowledge graph reasoning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 564–573.

[3] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[4] B. Shi and T. Weninger, "Open-world knowledge graph completion," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] W. H. O. of Science and T. Policy, "Call to action to the tech community on new machine readable covid-19 dataset," 2020.

[6] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, R. H. Zhang, W. Liu *et al.*, "Covid-19 literature knowledge graph construction and drug repurposing report generation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, 2021, pp. 66–77.

[7] C. Wise, M. R. Calvo, P. Bhatia, V. Ioannidis, G. Karypus, G. Price, X. Song, R. Brand, and N. Kulkani, "Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature," in *Workshop on Integrating Structured Knowledge and Neural Networks for NLP*, 2020, pp. 1–10.

[8] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu, "Drug repurposing for covid-19 via knowledge graph completion," *Journal of biomedical informatics*, vol. 115, p. 103696, 2021.

[9] M. Thilakaratne, K. Falkner, and T. Atapattu, "A systematic review on literature-based discovery: General overview, methodology, & statistical analysis," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–34, 2019.

[10] H.-T. Yang, J.-H. Ju, Y.-T. Wong, I. Shmulevich, and J.-H. Chiang, "Literature-based discovery of new candidates for drug repurposing," *Briefings in bioinformatics*, vol. 18, no. 3, pp. 488–497, 2017.

[11] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[12] G. Hahn-Powell, M. A. Valenzuela-Escárcega, and M. Surdeanu, "Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 103–108.