emerald PUBLISHING

Online Information F

The Unknown Knowns: A Graph-Based Approach for Temporal COVID-19 Literature Mining

Journal:	*Online Information Review*
Manuscript ID	OIR-12-2020-0562.R1
Manuscript Type:	Research Paper
Keywords:	COVID-19, Semantic Graphs, Natural Language Processing, Link Prediction, Machine Learning, Graph Temporal Analysis

SCHOLARONE™
Manuscripts

The Unknown Knowns: A Graph-Based Approach for Temporal COVID-19 Literature Mining

Abstract

Purpose: The COVID-19 pandemic has sparked a remarkable volume of research literature, and scientists are increasingly in need of intelligent tools to cut through the noise and uncover relevant research directions. As a response, we propose a novel framework. In this framework, we develop a novel weighted semantic graph model to compress the research studies efficiently. Also, we present two analyses on this graph to propose alternative ways to uncover additional aspects of COVID-19 research.

Design/methodology/approach: We construct the semantic graph using state-of-the-art Natural Language Processing (NLP) techniques on COVID-19 publication texts (>100,000 texts). Next, we conduct an evolutionary analysis to capture the changes in COVID-19 research across time. Finally, we apply a link prediction study to detect novel COVID-19 research directions that are so far undiscovered.

Findings: Findings reveal the success of the semantic graph in capturing scientific knowledge and its evolution. Meanwhile, the prediction experiments provide 79% accuracy on returning intelligible links, showing the reliability of the methods for predicting novel connections that could help scientists discover potential new directions.

Originality/value: To our knowledge, this is the first study to propose a holistic framework that includes encoding the scientific knowledge in a semantic graph, demonstrates an evolutionary examination of past and ongoing research, and offers scientists with tools to generate new hypotheses and research directions through predictive modeling and deep machine learning techniques.

Keywords: COVID-19, Semantic Graphs, Natural Language Processing, Link Prediction, Machine Learning

I. INTRODUCTION

In late 2019, COVID-19 started as a cluster of pneumonia cases in China, and has rapidly become a pandemic, resulting in hundreds of thousands of hospitalizations and deaths (Organization, 2020). Since then, scientists from all disciplines around the world have risen to the challenge and produced vast volumes of literature, aiming to understand the pandemic, its consequences, and come up with potential solutions. Allen Institute for AI in collaboration with well-known organizations systematically collected all literature containing COVID-19 and coronavirus-related research and made the collection named COVID-19 Open Research Dataset (CORD-19) publicly available (Wang et al., 2020b). This collection contains publications from peer-reviewed journals and pre-print services, and is updated regularly with the addition of new studies. Although having such a collection is helpful, the massive volume of studies it contains poses additional challenges to the researchers: the number of documents, especially those appearing on the preprint servers are more than anyone could read or share (Gog, 2020). One solution to scale-related data problems is to use graphical modeling techniques that provide a mechanism for exploiting and describing the complex data compactly and allows an effective utilization of the data (Koller and Friedman, 2009).

Graphical approaches can store data in the form of triplets that contain two entities (nodes/vertices) and the relation (edge/connection) between these entities. Such data representation techniques are particularly useful to link various relevant data sources together and to infer different knowledge aspects that are not possible to discover through texts manually (Project, 2020; Szekely et al., 2015). Several experts utilized graph-based literature mining techniques on

Online Information Review

CORD-19, mainly focusing on the easy querying aspect of the COVID-19 studies and made their graphs available. Table I has a summary of the current literature on COVID-19 graph-based approaches. While these tools are useful for researchers, they mine the literature in a query-able format, thus leaving the full potential of constructing and using graphs for COVID-19 research unexplored.

Table I ~~–~~ – Related work~~A survey~~ o~~f~~f graph-based COVID~~covid~~-19 literature mining studies and tools available, their characteristics, and the data sources used in their construction.

To better address these issues, we explore the potential of graph-based modeling using a novel research framework. First, we construct a novel, weighted, undirected semantic graph from the CORD-19 collection using only the studies published after the outbreak was acknowledged by WHO (January 2020) (Organization, 2020). This graph enables us to provide a new and compact summary of the COVID-19 research literature where the nodes contain scientific terms. In the algorithm we propose, these scientific entities and their relations are extracted from the full texts using a combination of state-of-the-art named entity recognition (NER) methods and our own methods that consider the existence of complex sentences with multiple scientific terms and relationship types. Taking these relations and NLP complexities into account enables this graph to capture the semantic relations between the entities present in the same sentences that are related based on meaning-wise properties. The connections in this graph are assigned weights based on an equation we propose that examines the text and dataset level occurrence statistics of the related entities. These weights indicate the association strength between the entities that serve as graph nodes. Following the construction of the semantic graph, we conduct two studies within the proposed research framework. Using the publication dates from those texts where the entities were extracted from, we conduct an evolutionary analysis on COVID-19 related terms. Next, we exploit the power of machine learning (ML) and link prediction methods on this graph to show the scientific community the possibility of intelligent forecasting of original research directions for COVID-19.

To our knowledge, this is the first study to propose a holistic framework that includes encoding the scientific knowledge in a semantic graph, demonstrates an evolutionary examination of the past and ongoing research, and offering scientists different tools to generate new hypotheses and research directions through predictive modeling and deep ML techniques. Also, by making these tools available, we are providing additional weapons necessary for battling with the current pandemic, and preparing for future pandemics. This novel framework is not only applicable for the COVID-19 related research, it is also applicable to any research problem where a set of publication texts are available for processing.

In the following sections, we explore the existing COVID-19 graphs, explain our methods for pre-processing the data, creating and exploiting the graph, report our results, and lay down our final remarks in the discussion and conclusion sections.

II. Materials and Methods

The proposed framework comprises of constructing a novel, weighted semantic graph from the selected dataset, and then using this graph in two distinct studies to best demonstrate its capabilities to further improve the COVID19 research knowledge: studying the evolution of knowledge across time, and predicting new research links that are so far undiscovered. In this section, the data collection, applied processing techniques, the algorithms used for constructing the semantic graph, and methods selected for evolution analysis and link prediction studies are explained in detail.

Online Information Review

A. Data and Pre-processing

As summarized in Table I, CORD-19 collection is vastly used by the COVID-19 research community. Made available by Allen Institute for AI with the collaborative efforts of the White House, Microsoft, AWS, bioRXiv, medRXiv, Kaggle, NIH/NLM, CSET, and Chan Zuckerberg Initiative (CZI), this collection contains publications from various sources (ArXiv, BioRxiv, MedRxiv, PMC, Elsevier, WHO, Medline) related to coronavirus and COVID19 research, is updated regularly and has been available to the public use1 (Wang et al., 2020b). The website also includes previous versions of this dataset. To construct our novel framework, we selected the latest version available at the beginning of our study (2020-05-19 version), containing 128,492 publications according to the meta-data file. Since then, newer versions became available. However, since the goal of this study is to propose a novel framework to researchers rather than directly returning results, we decided to refrain from constantly updating the collection. This should give more freedom to those who would want to reproduce the study on their own data collections.

Before performing NLP operations, we process the collection through a set of important data elimination, correction, and pre-processing steps. To capture only the COVID-19 related knowledge, we prioritize the elimination of the past articles from the collection. Although many publications in the collection did not contain publication dates, the dates were present for some papers in the meta-data file provided together with the collection. Also, there have been many issues with the meta-data file such as empty title cells and empty abstract cells, making it impossible to associate the meta-data to the publication texts which is the only way of associating dates and other meta-data to the complete texts. Therefore, various rule-based matching operations are conducted using the titles, abstracts, and authors recovered from the publication texts and from the meta-data cells. For those that can be matched, empty cells are filled with correct data. To further recover the remaining missing publication dates, we used Europe PMC RESTful Web Service2 to retrieve the publication dates by querying the digital object identifiers (DOI's). We also searched over the publication dates within the meta-data file to make sure the dates were accurate - and corrected those that did not match those retrieved via the Web Service. After all these efforts, the publications with remaining missing dates are eliminated. Also, those texts that appear both in archival websites and peer-reviewed journals are reduced into a single text. Next, texts containing less than five sentences are removed from the collection to make sure we only use the full-sized scientific publications.

After these text-size and missing date-related eliminations, 57,814 unique publication texts remained in the dataset. Papers were further filtered out based on language and dates. Following a language-detection and elimination of non-English studies, we selected "January 1st, 2020", the date the outbreak was acknowledged by WHO (Organization, 2020), as the starting date to separate COVID-19 studies from past coronavirus research. The remaining full texts are then pre-processed using standard NLP techniques: removing citations, elements within parentheses, and punctuation except for the end-of-sentence marks as well as lowercase conversion. The scripts developed for these processes are available from the project's GitHub directory[1].

1 ~~Temporarily removed for blind review~~ https://github.com/ubayram/COVIDGraphProject

Online Information Review

B. Constructing the COVID-19 Semantic Graph

Constructing a graph begins with the decisions of what nodes and their connections should represent. In this study, the aim is to create a graph that first captures scientific entities from the COVID-19 studies, including recently introduced terms. Secondly, semantic associations are included through weighted connections to serve as a compact COVID-19 knowledge summary. To accomplish this aim, we select to use state-of-the-art NER techniques together with our own algorithms.

NER is a crucial yet challenging task in literature mining. It allows the recognition and extraction of entities from the text. Deep learning (DL) models are increasingly used for providing scientifically-contextualized embedding models for NER (Zhao et al., 2020). To ensure high performance for the NER where novel scientific terms would be fully recognized, we experiment with two deep learning-based models: fine-tuned SciBERT (Beltagy et al., 2019), and scispaCy (Neumann et al., 2019). The fine-tuned SciBERT model presented inconsistencies in recognizing COVID-19 entities based on our observations, thus we focused on the ScispaCy model. In fact, ScispaCy provides specialized models that leverage the spaCy library (Honnibal and Montani, 2017), and contains models for performing NER on biomedical, scientific and clinical texts (Neumann et al., 2019). ScispaCy showed excellent performance on catching novel COVID-19 related terms during the initial experiments. Among the few available trained scispaCy models, we selected "en-core-sci-lg-0.2.4"4 for its performance during our initial experiments. However, off-the-shelf methods cannot promise high performance in the face of complex sentences with passive tense, multiple scientific terms, multiple verbs, objects, and subjects present. To overcome these obstacles, we develop the algorithm in Figure 1 for extracting the NER results, while taking advantage of the available tool, scispaCy.

Figure 1: The algorithm developed for the extraction of nodes and connections for the COVID-19 semantic graph using NLP techniques.

After collecting the entity pairs from texts, we construct the weighted, undirected semantic graph. Every unique entity is represented as a node, whether it contains a single word or a phrase. Since the entity pairs are returned from each sentence based on the existence of a semantic relation between them according to their roles in the sentences through POS-tagging, the constructed graph becomes a semantic graph. Subsequently, the semantic properties are further enriched by introducing weights on the connections where the weight values express the strength of the semantic relations, i.e., higher the value, stronger the relation. In the following paragraphs, we explain how the strengths of the relations between entities are captured using mathematical equations.

In graph theory, our graph can be defined as G(V, E) where V is the list of nodes (vertices) and E represents the weighted connections (edges). For an entity pair i = (a, b) with at least 10 relations returned after the algorithm is run over the complete collection, an undirected connection is created between a and b. In fact, entity pairs that occurred less than 10 times out of thousands of texts, millions of entity pairs are discarded. Such elimination is important to remove entities that have typos, therefore helps us to decrease the noise present in the graph. To compute the weight w_i for a connection i = (a, b), we developed Equation 1.

$$w_i = \frac{1}{N}\sum_{k=1}^{N} \frac{a_i^k}{\left(\sum_j^{n_k} a_j^k\right)} \qquad (1)$$

In the computation of weight w_i in Equation 1, N represents the number of texts remaining in the collection after all the previous eliminations (N = 11,066). For each text k, α_{ik} is the number of occurrences of the link i = (a, b), which is the number of occurrences of the entity pair a and b in the current text k. Collecting this statistic allows us to

Online Information Review

enumerate the importance of the pair within the text, e.g., the pairs repeated for many times in the text are very likely representative of the research terms discussed within that text. Next, this number is divided by the number of all links (connections) detected from the current text k, including the repeated entity pairs. This division eliminates the effect of a text's length over the occurrence of the entity pair (a, b), acting as a normalization factor that creates a balanced weight computation from texts of varying sizes. Aggregating all the collected occurrence ratios for the connection i provides the comprehensive collection-level connection strength. Finally, dividing the aggregated ratios by the size of the collection normalizes the strength into a weight value, thus returning the final semantic graph weights assigned to the connections in the graph. To keep the dynamic property of the graph of the semantic graph, we separately preserve the publication dates of the texts where each entity pair is detected from.

C. Evolution Analysis

Constructing semantic graphs from time-domain data has an advantage: it adds a dynamic property to these graphs. Properties of these graphs, such as the weight values, number of connections each entity has, etc., evolve through time with the addition of nodes and edges (Leskovec et al., 2007). Studying these dynamic changes has the potential of revealing interesting phenomena related to the data, while it may also provide cues to foreshadow the future (Bright et al., 2019). These potentials have led many researchers to study the evolution of various graphs, from brain networks (Thompson et al., 2017) to drug traffic networks (Bright et al., 2019). Therefore, to explore the dynamic potential of COVID-19 research progression across time domain, we explore what an evolution analysis can reveal.

Among various graph evolution analysis methods, we selected two approaches for the evolution analysis of COVID-19 research studies. One of the selected evolution analysis methods is volatility, which is the computation of the average connectivity in the graph across time (Thompson et al., 2017). Computing the connectivity of the graph over time returns a shape for the distribution of the volatility that can reveal interesting facts about the studied phenomena. The increase in connectivity over time reveals how many novelty connections are introduced to the graph (Thompson et al., 2017). The second selected evolution analysis method is called preferential attachment. Explained as the rich-get-richer phenomenon (Bright et al., 2019), computing and visualizing this across time demonstrates how some new entities in a semantic graph prefer to attach to highly connected entities in the graph. In the COVID-19 research domain, this method is expected to express how the research subject popularity evolved during this time period.

D. Link Prediction

Link prediction is the task of computationally deciding whether a pair of nodes in a graph should be connected based on their existing connections to other nodes (Liben-Nowell and Kleinberg, 2007; Martínez et al., 2016). This could particularly be interesting to further extend the future COVID-19 research directions. Thus, we apply a link prediction study on our semantic graph. We experiment with three different methods for evaluating how effective each method is on the task of link prediction over our semantic graph with the goal of selecting the best performing method for ultimate link prediction application.

Among the rich set of available link prediction methods, we are limited by those that consider the weighted and undirected properties of our graph. It is possible to implement a link prediction method that focuses only on the structural properties between the connections, however, implementing such a method on our semantic graph that comprises compact connections that are dense around the popular COVID-19 related terms is expected to return biased predictions, mainly around the highly connected nodes. Instead, we focus on selecting methods that consider the connection weights in their algorithms besides the other properties of the graph. For the first approach, we select a simple, well-known method called Jaccard coefficient (Liben-Nowell and Kleinberg, 2007) as a baseline link

Online Information Review

prediction method for the evaluation and for comparing the selected methods. For the next method, we select Adamic/Adar index (Adamic and Adar, 2003), which is also a known and popular method that is also simple. As our last approach, we select among the recent methods that take advantage of various complex algorithms including deep learning methods for providing link prediction such as GraphWave (Donnat et al., 2018), DeepWalk (Perozzi et al., 2014), and struc2vec (Ribeiro et al., 2017). One method that fulfills our goal is called node2vec, performs random walks within the graph and computes graph embeddings while taking the connection weights into account (Grover and Leskovec, 2016). In the following subsections, we explain the various steps undertaken for graph preparation, link prediction algorithms selection, and ML method implementation used in the experiments.

1) Sampling the graph for link prediction: Before performing link prediction over the semantic graph, we focus on evaluating how well the selected approaches fit the data. This evaluation process requires sampling a set of connections from the semantic graph. Utilizing the functions from the StellarGraph library (Data61, 2018), we randomly sample 50% of the connections from the semantic graph while preserving the graph connectedness. The resulting graph, G_i contains the same number of nodes as the original graph G but has the sampled edges removed. This graph G_i is used for processes such as obtaining node2vec embeddings from the graph. Sampled connections represent 50% of the actual links in the semantic graph and are considered to be positive samples. Negative samples, on the other hand, represent the absence of connections, and include a set of links with the same size as positive samples. Next, the sampled edges are further divided into training, validation and test sets. For this division, we select the rates as 50% for training, 15% for validation, 35% for test sets.

We repeat the edge sampling and training-test-validation division five times under Monte Carlo cross-validation framework (Dubitzky et al., 2007). At each run, we sample distinct sets of connections from the graph, getting different embedding results and performing link prediction on different data combinations from the randomly split training-test-validation sets, achieving separate performance evaluations. This experimentation pipeline enables us to test the stability of the link prediction and ML results. An unstable method for instance would return a high result from one sample set or experiment and low in another one.

2) Weighted Jaccard coefficient: Influenced by the weighted Jaccard distance equation (Ioffe, 2010), we construct the following equation to return a weighted Jaccard coefficient value J(u, v) for a potential link between the unconnected nodes u and v:

$$J(u,v) = \frac{\sum_{y \in \{N(u) \cap N(v)\}} w(y)}{\sum_{x \in \{N(u) \cup N(v)\}} w(x)} \qquad (2)$$

Similar to the traditional Jaccard coefficient, which is the number of common neighbors between u and v divided by the number of nodes in the union of their neighbors (Liben-Nowell and Kleinberg, 2007), we compute the common neighbors for the numerator and the union for the denominator. In contrast to the traditional method, we introduce weighted degrees into the equation. A weighted degree of a node can be explained as the sum of all the weighted direct connections it has to its neighbors. For each node in the set of intersection or union of neighbors, we compute the weighted degrees and then sum over all the nodes in the set. Through this change in the equation, we get to include the weights of the semantic graph into account that is straightforward to interpret in terms of link prediction: the node pairs that have many common neighbors with strong connections (high weights) are more likely to be connected to each other, thus having high J(u, v) scores.

3) Weighted Adamic-Adar index: The second link prediction selected is the weighted version of the Adamic/Adar index. In the standard Adamic/Adar index, the score of connectivity between the two unconnected nodes u and v is

Online Information Review

computed by finding the number of their common neighbors, and summing the inverse logarithms of these neighboring nodes' degrees (Adamic and Adar, 2003). In our implementation, we introduce the weighted degrees to the equation as follows:

$$S(u,v) = \sum_{y \in \{N(u) \cap N(v)\}} \frac{1}{-log\ (w(y))} \qquad (3)$$

Since the weights in our semantic graph were normalized into the range [0, 1], we multiply the results with −1 to avoid getting negative scores due the logarithm operation. Thereby, it is important to realize that these scores behave differently from the standard Adamic/Adar index where a high number of common neighbors would return low scores in the standard approach. In this equation, those nodes u, v with high number of common neighbors, large weighted degrees will return high scores similar to the weighted Jaccard coefficient. So, the pairs with high S(u, v) score are those with many common neighbors with strong connections to the rest of the graph.

4) Link prediction with node2vec: Influenced by the recent word2vec word embedding method, node2vec learns to map the nodes to a low-dimensional space of features that maximizes the possibility of preserving the structural properties of the graph (Grover and Leskovec, 2016). The method performs second order, biased random walks over the graph to sample the neighborhoods, to learn which nodes commonly appear together or are in similar neighborhoods in low complexity while taking the connection weights into account. It makes use of two different strategies for each source node k: breadth-first sampling (BFS) and depth-first sampling (DFS). BFS restricts the sampling at immediate neighbors of the source node, while DFS sequentially samples at increasing distances from the source. The node2vec method interpolates between the BFS and DFS methods within the random walks. In this study, the StellarGraph implementation of node2vec is used (Data61, 2018).

As expected of complex algorithms, this method also has certain disadvantages. Specifically, the number of parameters to be set by the user is excessive, and the computational time requirements of the approach makes it near impossible to implement a grid-search operation for finding the best parameters for the data. The key problem with these parametric ML methods is that selecting incorrect parameters may have a significant effect on the performance (Domingos, 2012). In this study, we selected to work with a set of parameters that are used in the original node2vec study (Grover and Leskovec, 2016), while tuning down some of them to speed up the computational time. Among these parameters are the dimension of the embedding features (= 128), the random walk length per source node (= 60), the window of the number of nodes (= 12), the number of walks to perform (= 10), and the p and q which represent the BFS and DFS parameters. To find the effect of p and q on the performance evaluations over the validation set data, we conducted a grid-search over the range {0.25, 1, 4} for both. However, we did not observe a significant impact on the results, so we selected the value 0.25 for both and performed the experiments under these settings. We also performed another selection for the remaining word2vec settings: the minimum count as 0 to avoid frequency-based elimination, and the skip-gram model for its popularity on NLP studies. We kept the rest of the parameters at their defaults in the gensim implementation (Řehůřek and Sojka, 2010).

The node2vec method returns node embeddings. For the ML task of predicting whether two nodes should be connected, we need to convert these node embeddings into feature values for connections between the nodes. Therefore, we experiment with three binary operators out of the four that were used in the original node2vec study to compute the distance between these embeddings to return a feature value: L1, Hadamard, and average (Grover and Leskovec, 2016). To determine the best operator for the feature value extraction, we test the trained classifier over the validation samples and use the best performing operator for the actual tests. Finally, we apply standard scaling that removes the mean and scales the features to unit variance to obtain ideal ML performance.

Online Information Review

5) Evaluation with machine learning: The link prediction methods return a score for every node pair they are given. To convert a score into a decision of connecting or not connecting these nodes requires the aid of ML methods. Due to the inherent advantages and shortcomings of each ML algorithm, and depending on the data, we take the possibility of simple ML methods performing as well as, or even better than the complex ones into account (Bayram et al., 2019). We thereby select three supervised methods that have different methodologies and complexities to test the effect of ML method selection over the link prediction results: a simple classifier called naïve Bayes (Buitinck et al., 2013), another method called logistic regression that requires parameter tuning to learn how to separate the training data into the two classes (class 1 = connect, class 2 = not connect) through cross-validation (10-fold with maximum 2,000 iterations) (Buitinck et al., 2013), and a deep learning classifier. In the last classification technique, a multi-layer perceptron, our aim is to test whether the increased complexity of the ML method would improve the link prediction results. Using the Tensorflow API (Abadi et al., 2015), we construct a feed-forward neural network with four layers excluding the input layer, with 100 neurons in the first, 50 neurons in the second, 10 neurons in the third, and 2 neurons at the final output layer. Without the presence of a dropout, all layers are connected densely. Within the hidden layer neurons, we select the hyperbolic tangent "tanh" as the decision function, meanwhile, for the output layer neurons we select the "sigmoid" to get probabilities for class 1 and class 2, i.e., returning a link to connect the two nodes or not. Optimization is done by Adam optimizer with sparse categorical cross-entropy loss function for the selection of two output layer neurons that represent the two classes. As overfitting a model is one of the major issues in ML (Domingos, 2012), we apply an early stopping scheme using the validation set classification accuracy during the training with the patience of 25 epochs. As the batch size, we select 2,500, as the data collection comprising node pairs is large.

III. Results

The novel COVID-19 semantic graph constructed using the CORD-19 collection returns a compact representation of interconnected scientific terms within 265,526 nodes, and 1,075,627 connections. As the large number of connections shows, this is a densely connected graph. In our evaluations, however, we did not detect small-worldness in the graph as commonly demonstrated in many big graphs (Milgram, 1967). As part of the proposed framework, we perform the experiments explained in the previous section to first analyze the evolution of research across time. Next, we perform link prediction experiments to demonstrate the potential of finding new research links using our framework.

A. Evolution of COVID-19 Research

Before evaluating the evolution analysis, we plot a histogram of publications used in the graph construction as a baseline for interpreting the time-domain evolution results. The histogram in Figure 2 shows that the majority of the publications are from April and May, while late May and early June had very few publications. Considering the CORD-19 version used in this study, some of the publication dates can be ahead of the actual date, thus explaining the appearance of June articles in the collection.

Figure 2. Histogram of the publication dates of the research studies used in construction of our semantic graph.

Online Information Review

The initial evolution analysis focuses on volatility, which is a measure of global graph connectivity. Figure 3 shows the average connectivity computed as the average number of new connections introduced per node every day. Although there is a slight exponential increase in overall connectivity, there is no drastic change as the addition of new connections to the graph seems to be aligned well with the introduction of new nodes. Figure 4 shows the same growth in another form; the expansion of the graph as the number of new connections and new nodes added to the graph bi-weekly. In this figure, the overall plot also appears in a linear form. Looking closely at the data points in this figure shows the sporadic growth trend. Four data points from January and February being close to each other reveal that between these two months, a similar number of new scientific terms or term associations were present in the publications from these dates. This can be explained by Figure 2 where it is shown that the number of publications from these months were not as high as the following months. However, the data points from March and April have huge jumps in the plot, observed in sizable gaps between these points. This can partially be explained by the large numbers of new publications from these months (cf. Figure 2), while gaps show that these publications had many new scientific terms and new connections introduced to the graph in proportional amounts compared to the past months.

Figure 3. Evolution of average connectivity in the semantic COVID-19 graph computed as the average number of new connections introduced per node every day.

Figure 4. Evolution of average connectivity in the semantic graph through time, demonstrated as the growth of the graph computed bi-weekly.

Following the simple global evolution analysis, we focus on local evolution of some popular terms in the graph. For the preferential attachment analysis, popularity is selected as the primary criteria for preferentiality, which is the rich-get-richer ideology. Popularity in this domain refers to the degree of a node, i.e., higher the degree, the more popular a node is. From Figure 5 to Figure 7, the popularity evolution of the most popular terms in the semantic graph are visualized, where the groups are based on their contextual similarity so the visual analysis can reveal useful clues on COVID-19 research progress. Concurrently, the homophily amongst the groups of terms will promote deeper questions about the contexts being discovered.

Figure 5. Popularity evolution of the COVID-19 related ``condition'' terms in the semantic graph nodes. For better visualization, the set of selected terms are divided into two figures, and y-axes are converted into logarithmic domain.

In Figure 5, “pneumonia” is the most popular term followed by “sepsis” and “acute respiratory syndrome” since the beginning of the pandemic research. “Coronaviruses” falls under the broader category of respiratory disorder viruses affecting the lungs (Kuba et al., 2005), hence we were not surprised to find many terms associated with infection and respiratory disorders as the most popular terms. Pneumonia and sepsis/septic shock can be caused as a result of viral infection in the body (Lin et al., 2018). Acute respiratory syndrome, on the other hand, occurs due to inflammatory response of the immune system to fight against the infections (Lin et al., 2018). Predictably, the “lung” node was the most popular from the start of the pandemic research among the human body group.

Online Information Review

Figure 6. Popularity evolution of the immunity, hospital, drug-terms, and treatment-related terms as semantic graph nodes.

Since viral infections lead to inflammation as an immune response to protect the body against the infection (Lin et al., 2018), it is imperative to understand how the inflammation/immunity related nodes evolve. In Figure 6, all of the terms are part of ongoing COVID-19 research since the initial time point of the pandemic. "Antibody" is the most popular term followed by "cytokine storm" and "immune response". Antibodies are an important immune reaction of the host to combat infections. Cytokine storm thus sets in as an immune response.

Between hospital/ppe related nodes in Figure 6, interestingly, "ICU" and "intensive care unit" were the most popular until mid-April. After that, "ppe" and "personal protective equipment" nodes became more popular. This is likely caused by an increasing general awareness of the issues of hand hygiene, ppe, etc. Among the drugs that are used to treat viral-infected patients, in Figure 6 "antibiotic" turned out to be the most popularly researched nodes since the beginning. "Chloroquine" and "hydroxychloroquine", two proven antimalarial drugs (Pastick et al., 2020) started becoming popular around late March once they showed promise in treating coronavirus patients. "Remdesivir", the drug currently in clinical trials for showing promising effects against SARS-Cov-2 (Beigel et al., 2020) slowly gained popularity since the beginning of March. One note on the drug-related nodes is that we exclude the term "antivirus" as it is difficult to distinguish its drug-specific use-cases from the generic mentions. Within the treatment/diagnosis terms in Figure 6, "vaccination" and "mechanical ventilation" were the most popular. To curb a pandemic of such scale, vaccination is the most important tool. However, until we have a vaccine, severe patients need to be helped using mechanical ventilation.

Figure 7. ~~P~~Popularity evolution of the viral/bacterial, device-related, social, and comorbidity-related terms in the semantic graph nodes.

Figure 7 shows that "pathogen" and "viral load" terms were popular until early April, after which the term "pathogen" became more popular in the graph in terms of new connections. Surprisingly, "ace2" did not show up to be a popular node, although a lot of research has been carried out with it. Ace2 protein is found on the viral spike-protein and viruses enter the human cell through ace2 receptors in the host body (Ciaglia et al., 2020). We presume this is because we could only select the most mentioned terms and could not analyze similar terms (ace2/ace2 receptor/angiotensin converting enzyme) together. Foreseeably, among the device terms in Figure 7, "intubation" was the most popularly researched node to aid in mechanical ventilation. On the other hand, node "telemedicine" started to become popular from late March when many countries started to take lock-down measures.

According to the Figure 7, amongst other diseases (comorbidities) or vulnerable population groups affecting the viral infection spread, "pregnancy" seemed to be the most popular since late February. "Cancer" research related to COVID has gone up exponentially since the beginning of the pandemic, while "hypertension" started to gain popularity in late March. Furthermore, research studies also considered the social aspect of the COVID-19 pandemic, mentioning many social terms as seen in Figure 7. "Quarantine" started getting popular in late January and "social-distancing" became popular in late March which matches the timeline of lock-downs in several countries. Though "aerosol" research started in mid-January, it did not seem to grow popular over time. Although most respiratory viruses can spread through aerosols, WHO is yet to ascertain the nature of the spread of the virus. Such information is very important to curb further spread of communicable diseases.

Online Information Review

B. Evaluating the Link Prediction on Our COVID-19 Semantic Graph

To ultimately provide predicted links connecting novel research areas for new COVID-19 research paths, we first evaluated three different link prediction methods: weighted Jaccard coefficient, weighted Adamic/Adar index, and node2vec to choose the best performing method. Since we decide on connecting entities based on link prediction scores, we experiment with three supervised ML methods: naïve Bayes, logistic regression, and multi-layer perceptron. Each link prediction and ML method combination is a completely separate experimental scenario, and thus might return different results. Also, due to the randomness factors present within the methods such as node2vec and multi-layer perceptron, different runs of the same combination on the same data might return different results. Alternative samples may also return completely different results for the same combination as well. To capture the range of such differences, we perform 5-fold Monte Carlo cross-validation and return the mean and standard deviation of the performance scores. The returned classification scores range between [0, 1], where a classification score close to 1 would mean the nodes should be connected. Therefore, we perform the classification by thresholding the scores by 0.5, where the values equal to or above this threshold are labeled as "connected" and those below this value are labeled as "unconnected". Comparing these predicted labels to the actual hidden labels enabled us to return the scores seen in Table II.

Table II - Average accuracies and standard deviations of the link prediction results classified over the test set and 5-fold cross-validation, and the significance of the difference between the two methods with p-values.

The results in Table II reveal several interesting observations about our semantic graph and the implemented methods. For example, all the results are ranging between 68% and 79% average accuracies with low standard deviations expressing the success of the graph in modeling the relations among the scientific terms. The consistency of results as outlined in the table supports the success of our semantic graph construction.

Different combinations of link prediction and ML methods also point out various fascinating phenomena. For example, link prediction with weighted Jaccard coefficient returns the best result when it is combined with the multi-layer perceptron, while its combination with the other simple classifiers returns the lowest results (cf. Table II). This can be explained by the success of the multi-layer perceptron in capturing the relations between the J(u, v) scores and the actual connectivity status of the nodes (u, v). The second method, weighted Adamic/Adar index returns consistently top results per classifier. This result highlights the superiority of this method over Jaccard coefficient, which further emphasizes the importance of experimenting with different methods, since data can bring about unforeseeable behaviors.

The results of the complex link prediction method, node2vec, also reveal compelling observations. For example, compared to the weighted Adamic/Adar index results, node2vec results per classifier are almost never better than the simpler link prediction method. It shows that complexity does not necessarily guarantee better results. This can also be seen in the original node2vec study (Grover and Leskovec, 2016), where among the different datasets experimented on, Adamic/Adar appears to perform as good as the node2vec method under some scenarios. Meanwhile, node2vec combined with multi-layer perceptron succeeds to marginally outperform the simple methods on our link prediction tests. An interesting observation from this combination is the magnitude of the Area Under the ROC curve (AUC) scores. While the other experiments would return ~80% AUC scores, the combination of node2vec and multilayer perceptron returns above 90% AUC. This could indicate that performing the link prediction classification by selecting

Online Information Review

those with classification score above 0.5 as the positive, and below 0.5 as negative may not be a good classification scheme for this specific combination. For example, selecting a different threshold for classification could yield higher classification accuracies for this combination. Therefore, for returning actual links predicted from our semantic graph, we select this combination of methods.

C. Exploring Predicted Links for COVID-19 Research

As the final stage of our framework, we present actual link predictions for the unconnected entity pairs from the semantic graph using the node2vec with multi-layer perceptron algorithm. To perform this experiment, we work on one of the sampled sets from the previous experiments since the multi-layer perceptron requires training and validation sets for its learning and tuning processes. Therefore, we select one of the 5-fold Monte Carlo cross-validation samples that contains a graph with 50% of the actual connections from the complete semantic graph, while the remaining links are sampled for the training-validation and the previous test set. In this setting, we compare the unconnected entity pairs from the small graph to the connections in the overall semantic graph to make sure we only collect entity pairs that are unconnected in the actual semantic graph. The remaining operations are the same as the past experiments where we compute the node2vec embeddings from the sampled graph, collect training, validation, and the new test set embeddings, find the best fitting operator for converting the embeddings into feature values, and proceed with obtaining the classification scores for the actually unconnected entity pairs.

Table III - Some of the predicted links from the half of the semantic graph, from actual unconnected pairs of entities seen below are obtained from the node2vec graph embedding method combined with multi-layer perceptron.

The collection of sampled pairs is large. To demonstrate an interpretable subset of the link prediction results, we provide some top scoring, unconnected entity pairs and their classification score in Table III. Although not connected in the graph, our link prediction algorithm returned several entities that should be connected. While some are generic such as the predicted links between “pbst buffer” and “reaction buffer”, or “nucleotide identity matrix” and “pwm2 m”, many of them are biologically specific terms such as “sars-cov binding” and “nonsmall cell lung cancer”, or “effector t cell activation” and “rituxan”. Such predictions are also supported by the literature. For example, “rituxan” has been linked to “effector t cells” in many studies (Neel et al., 2019). Our novel framework thus can predict many important otherwise unconnected relations.

IV. Discussion

In this study, we proposed a novel research framework that includes construction of a novel, weighted semantic graph from the CORD-19 collection of publications, and two applications to help the COVID-19 research community in their current and future efforts. To demonstrate the viability of our framework, we selected the 2020-05-19 version of the collection which was the newest version at the time of our study. The framework remains applicable on more updated versions of the dataset and its corresponding scripts are available at the project’s Github page.

We further presented two useful downstream analyses over our semantic graph: evolution analysis and link prediction. Evolution analysis helps to understand how the pandemic and associated terms have changed over time. For example, in the beginning of the outbreak patients were put on antibiotic and antiviral drugs, but antimalarial drugs like hydroxychloroquine and chloroquine became popular later on, a finding also observed in our study. Link

Online Information Review

prediction on the other hand is an excellent way to bridge a connection between unseen entities that have newly emerged and seen entities already present. For example, links between novel vaccines under the coronavirus study and unseen entities newly emerged can be gauged using link prediction.

Just like any research study, we also faced struggles and shortcomings during the implementation. One of the hurdles was handling the computational complexity of the node2vec link prediction algorithm. This parametric method requires many parameter settings, and with the fear of default parameters not fitting the data, we attempted to implement grid search operations to auto-select the parameters using the validation set results. However, as the graph size was large even after sampling half of the connections, the computational time was very constraining even while using additional resources like AWS servers. These shortcomings restricted us to manually and selectively set these parameters by relying on the defaults given in the original study. However, there could be a parameter combination that would perform much better link prediction than those reported in Table II and that will remain unexplored until grid-search is conducted. Overall, we cannot claim that Adamic/Adar is definitely better than node2vec due to the lack of tuning, but considering the complexity differences and obtained results, we can highlight that simple methods could be just as efficient under the circumstances.

For researchers that would like to use the framework or some of the proposed methods presented in this study, we suggest to keep some considerations in mind. For example, in case a researcher implements this study on a more recent version of the CORD-19 collection, using all of the publications would yield a much larger semantic graph that requires large computational resources. Therefore, they should consider reducing the graph size whether by not using the full texts but focusing on certain sections only for NER applications, or eliminating irrelevant publications from their datasets. Alternatively, they could choose to adapt the framework to work with parallel graph engines that could handle processing large graphs. Regardless of the decision, implementation of our semantic graph will remain more efficient than any other approach requiring the processing of tens of thousands of publications involving manual labor. Another suggestion would be to change the sample size fractions used in our experiments for more efficient application of complex methods such as node2vec. The evolution analysis study can be further extended in future works based on specific research goals, while additional link prediction methods in collaboration with different ML classifiers can be introduced in the future studies as well. Furthermore, the proposed framework can be generalized to scientific literature emanating from various research fields, with very few modifications such as using a domain-specific named entity recognition model.

V. Conclusion

The COVID-19 pandemic has sparked a remarkable volume of research literature, and scientists are increasingly in need of intelligent tools to help them eliminate the noise and uncover the relevant information. To answer these needs, we proposed a novel framework comprising three main parts: the construction of a novel weighted semantic graph model to compress the research studies, evolution analysis of COVID-19 research from the graph, and prediction of new research directions using machine intelligence on the graph. All the experimental findings revealed the success of the semantic graph in capturing scientific knowledge and its evolution. Meanwhile, the prediction experiments provided high accuracy, showing the high potential of such applications to return novel research directions. The predicted links from the graph are interpretable, and can be of value for future research. To our knowledge, this is the first study to propose such a holistic framework that includes an encoding of the scientific knowledge in a semantic graph format while demonstrating strong capabilities one can achieve from this framework and the proposed techniques. The tools we are offering would help scientists to generate new hypotheses and research directions.

Online Information Review

Making these tools available gives them additional weapons for battling against the current pandemic, and also for tackling future research problems.

REFERENCES

(Abadi *et al*., 2015) Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D.,´ Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale´ machine learning on heterogeneous systems. Software available from tensorflow.org.

(Adamic and Adar, 2003) Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

(Babcock *et al*., 2020) Babcock, S., Cowell, L. G., Beverley, J., and Smith, B. (2020). The infectious disease ontology in the age of covid-19. In *OSF Preprints*. Center for Open Science.

(Bayram *et al*., 2019) Bayram, U., Pestian, J., Santel, D., and Minai, A. A. (2019). What's in a word? detecting partisan affiliation from word use in congressional speeches. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

(Beigel *et al*., 2020) Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A., Kline, S., *et al*. (2020). Remdesivir for the treatment of covid-19—preliminary report. *New England Journal of Medicine*.

(Beltagy *et al*., 2019) Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

(Bras *et al*., 2020) Bras, P. L., Gharavi, A., Robb, D. A., Vidal, A. F., Padilla, S., and Chantler, M. J. (2020). Visualising covid-19 research.

(Bright *et al*., 2019) Bright, D., Koskinen, J., and Malm, A. (2019). Illicit network dynamics: The formation and evolution of a drug trafficking network. *Journal of Quantitative Criminology*, 35(2):237–258.

(Buitinck *et al*., 2013) Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108– 122.

(Ciaglia *et al*., 2020) Ciaglia, E., Vecchione, C., and Puca, A. (2020). Covid-19 infection and the predictive ace2 soluble levels: the favourable protection of children and women. *Front Pediatr*, 23(8):206.

(Collabovid, 2020) Collabovid (2020). Exploration of covid-19 publications. https://collabovid.org/. Accessed: 2020-07-31.

(Data61, 2018) Data61, C. (2018). Stellargraph machine learning library. https://github.com/stellargraph/stellargraph.

(Domingo-Fernandez *et al*., 2020) Domingo-Fernandez, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., *et al*. (2020). Covid-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of covid-19 pathophysiology. *BioRxiv*.

(Domingos, 2012) Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

(Donnat *et al*., 2018) Donnat, C., Zitnik, M., Hallac, D., and Leskovec, J. (2018). Learning structural node embeddings via diffusion wavelets. In *International ACM Conference on Knowledge Discovery and Data Mining (KDD)*, volume 24.

(Dubitzky *et al*., 2007) Dubitzky, W., Granzow, M., and Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

(Floros *et al*., 2020) Floros, D., Pitsianis, N., and Sun, X. (2020). Lg-covid19-hotp: Literature graph of scholarly articles relevant to covid-19 study. https://lg-covid-19-hotp.cs.duke.edu/. Accessed: 2020-05-20.

(for AI, 2020) for AI, A. I. (2020). Scisight. Retrieved from https://scisight.apps.allenai.org/. Accessed 2020-05-25. 10.5281/zenodo.3715505.

(Gog, 2020) Gog, J. R. (2020). How you can help with covid-19 modelling. *Nature Reviews Physics*, pages 1–2.

(Grover and Leskovec, 2016) Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

(Honnibal and Montani, 2017) Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

(IBM, 2020) IBM (2020). Ibm covid19 deep search. https://www.research.ibm.com/covid19/deep-search/. Accessed: 2020-07-31.

(Ioffe, 2010) Ioffe, S. (2010). Improved consistent sampling, weighted min-hash and l1 sketching. In *2010 IEEE International Conference on Data Mining*, pages 246–255. IEEE.

(Koller and Friedman, 2009) Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

(Kuba *et al*., 2005) Kuba, K., Imai, Y., Rao, S., Gao, H., Guo, F., Guan, B., Huan, Y., Yang, P., Zhang, Y., Deng, W., *et al*. (2005). A crucial role of angiotensin converting enzyme 2 (ace2) in sars coronavirus–induced lung injury. *Nature medicine*, 11(8):875–879.

Online Information Review

(Leskovec *et al*., 2007) Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es.

(Liben-Nowell and Kleinberg, 2007) Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

(Lin *et al*., 2018) Lin, G.-L., McGinley, J. P., Drysdale, S. B., and Pollard, A. J. (2018). Epidemiology and immune pathogenesis of viral sepsis. *Frontiers in immunology*, 9:2147.

(Martínez *et al*., 2016) Martínez, V., Berzal, F., and Cubero, J.-C. (2016). A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, 49(4):1–33.

(Milgram, 1967) Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

(Neel *et al*., 2019) N´ eel, A., Bucchia, M., Neel, M., Tilly, G., Caristan, A., Yap, M., Rimbert, M., Bruneau, S., Cadoux, M., Agard, C., *et al*. (2019). Dampening of cd8+ t cell response by b cell depletion therapy in antineutrophil cytoplasmic antibody–associated vasculitis. *Arthritis & Rheumatology*, 71(4):641–650.

(Neumann *et al*., 2019) Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

(Organization, 2020) Organization, W. H. (2020). Who timeline - covid-19. https://www.who.int/news-room/detail/ 27-04-2020-who-timeline---covid-19. Accessed: 2020-05-03.

(ORKG, 2020) ORKG (2020). Open research knowledge graph. https://www.orkg.org/orkg/. Accessed: 2020-07-31.

(Pastick *et al*., 2020) Pastick, K. A., Okafor, E. C., Wang, F., Lofgren, S. M., Skipper, C. P., Nicol, M. R., Pullen, M. F., Rajasingham, R., McDonald, E. G., Lee, T. C., *et al*. (2020). Hydroxychloroquine and chloroquine for treatment of sars-cov-2 (covid-19). *Open Forum Infectious Diseases*, 7(4).

(Perozzi *et al*., 2014) Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

(Project, 2020) Project, T. C. (2020). Covid*graph. https://covidgraph.org/. Accessed: 2020-05-20.

(Řehůřek and Sojka, 2010) Řehůřek , R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/ 884893/en.

(Ribeiro *et al*., 2017) Ribeiro, L. F., Saverese, P. H., and Figueiredo, D. R. (2017). struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 385–394.

(SilkData, 2020) SilkData (2020). Semantic maps. https://www.silkdata.ai/news/semantic-map-covid-news-article. Accessed: 2020-05-20.

(Szekely *et al*., 2015) Szekely, P., Knoblock, C. A., Slepicka, J., Philpot, A., Singh, A., Yin, C., Kapoor, D., Natarajan, P., Marcu, D., Knight, K., *et al*. (2015). Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*, pages 205–221. Springer.

(Thompson *et al*., 2017) Thompson, W. H., Brantefors, P., and Fransson, P. (2017). From static to temporal network theory: Applications to functional brain connectivity. *Network Neuroscience*, 1(2):69–99.

(UCSB, 2020) UCSB (2020). Graphdb visual graph. https://covid.geog.ucsb.edu/. Accessed: 2020-05-20.

(Vermeij, 2020) Vermeij, A. (2020). Covid-19 open research map. https://covid-19.kenedict.com/. Accessed: 2020-08-24.

(Wang *et al*., 2020a) Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., and Kanakia, A. (2020a). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

(Wang *et al*., 2020b) Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R. M., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D. M., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020b). Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706.

(Wang *et al*., 2020c) Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., Chauhan, A., Guan, Y., Li, B., Li, R., Song, X., Ji, H., Han, J., Chang, S.-F., Pustejovsky, J., Rah, J., Liem, D., Elsayed, A., Palmer, M., Voss, C., Schneider, C., and Onyshkevych, B. (2020c). Covid-19 literature knowledge graph construction and drug repurposing report generation.

(Zhao *et al*., 2020) Zhao, S., Su, C., Lu, Z., and Wang, F. (2020). Recent advances in biomedical literature mining. *Briefings in Bioinformatics*.

Online Information Review

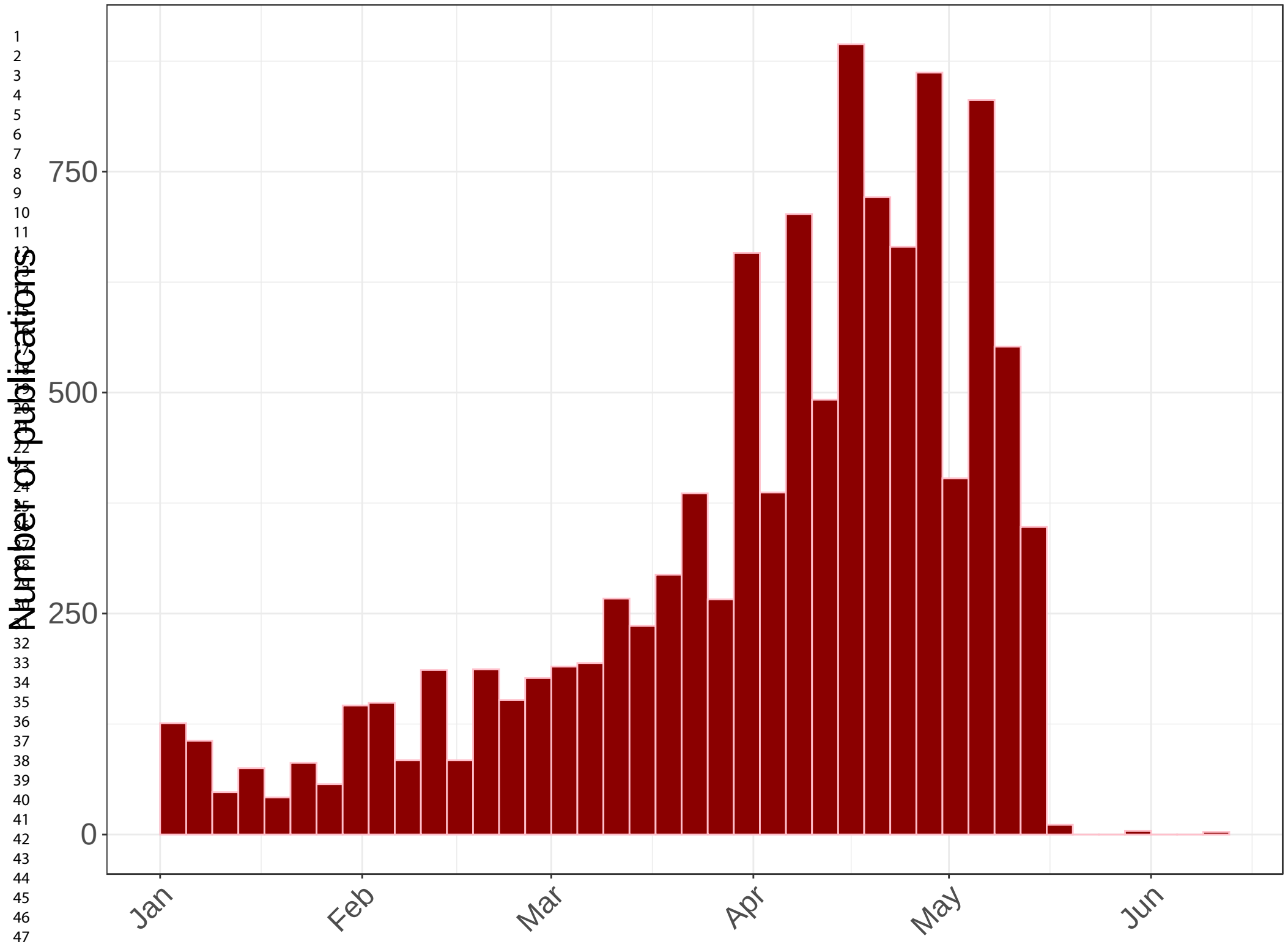
Study	Data	Type	Characteristics
COVIDGRAPH (Project, 2020)	CORD-19, Gene Ontology Resource, Ensemble Genome Browser, Lens COVID-19, Data Repository by John Hopkins Experimental Data from studies NCBI Gene Database	Web-based	Search for treatment options and Existing patents for users
COVID-19 Open Research Map (Vermeij, 2020)	CORD-19, Arxiv, Medrxiv, Biorxiv	Web-based	Relation between document abstracts and publication journal
GraphDB Visual Graph (UCSB, 2020)	Data Repository of Johns Hopkins University Manually collected	Web-based	Implication of the pandemics
Microsoft Academic Graph (Wang *et al*., 2020a)	Bing search results, CORD-19, PubMed	Web-based	Knowledge base of publications
LG-COVID19-HOTP (Floros *et al*., 2020)	CORD-19	Web-based	Scholarly articles with in-and-out citations self navigation
The semantic Map (SilkData, 2020)	CORD-19	Web-based	Information to researchers with relation among nodes
Scisight (for AI, 2020)	CORD-19	Web-based	Interactions between biological terms
Collabovid (Collabovid, 2020)	BioRxiv, MedRxiv, Pubmed, Elsevier	Web-based	Extraction of specific researches
IBM KG (IBM, 2020)	CORD-19, DrugBank, GenBank	Web-based	Deep searching with complex queries
ORKG (ORKG, 2020)	Manually curated	Web-based	Viral replication rate
Domingo *et.* al. (Domingo-Fernandez *et al*., 2020)	PubMed, Europe PMC, Targeting 2019-nCoV, LitCovid, University of Luxembourg dataset	Pre-print	Human-viral interaction, potential drug targets
DRKG (Wang et al., 2020c)	Manually collected peer-reviewed articles	Pre-print	Drug selection to curb pandemic
Bras et. al. (Bras *et al*., 2020)	Dimensions' COVID-19 and CORD-19 datasets	Pre-print	Trend analysis for quick data discovery
IDO-COVID-19 (Babcock *et al*., 2020)	Surveillance data	Pre-print	An ontology to help in data integration and sharing

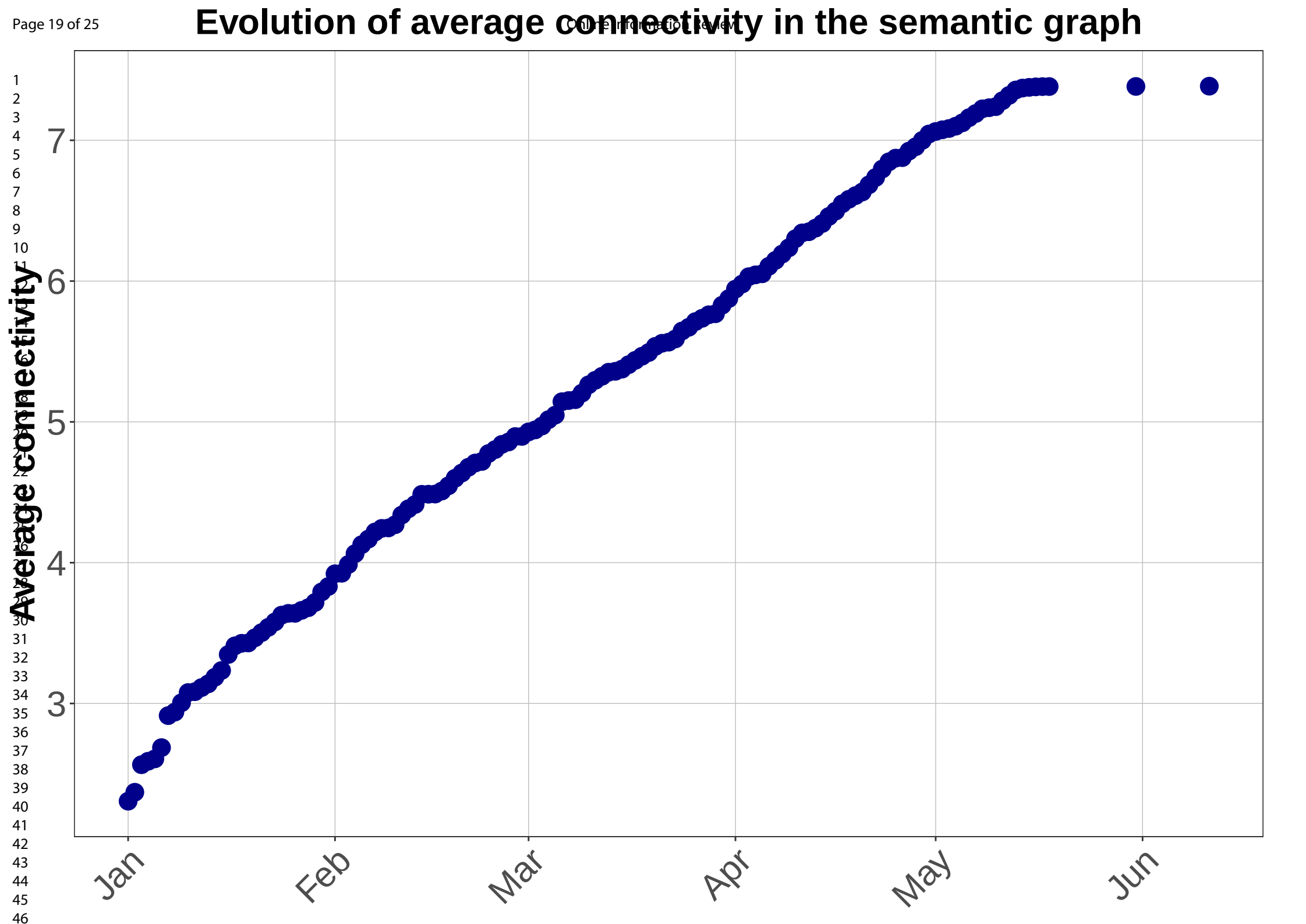
Online Information Review

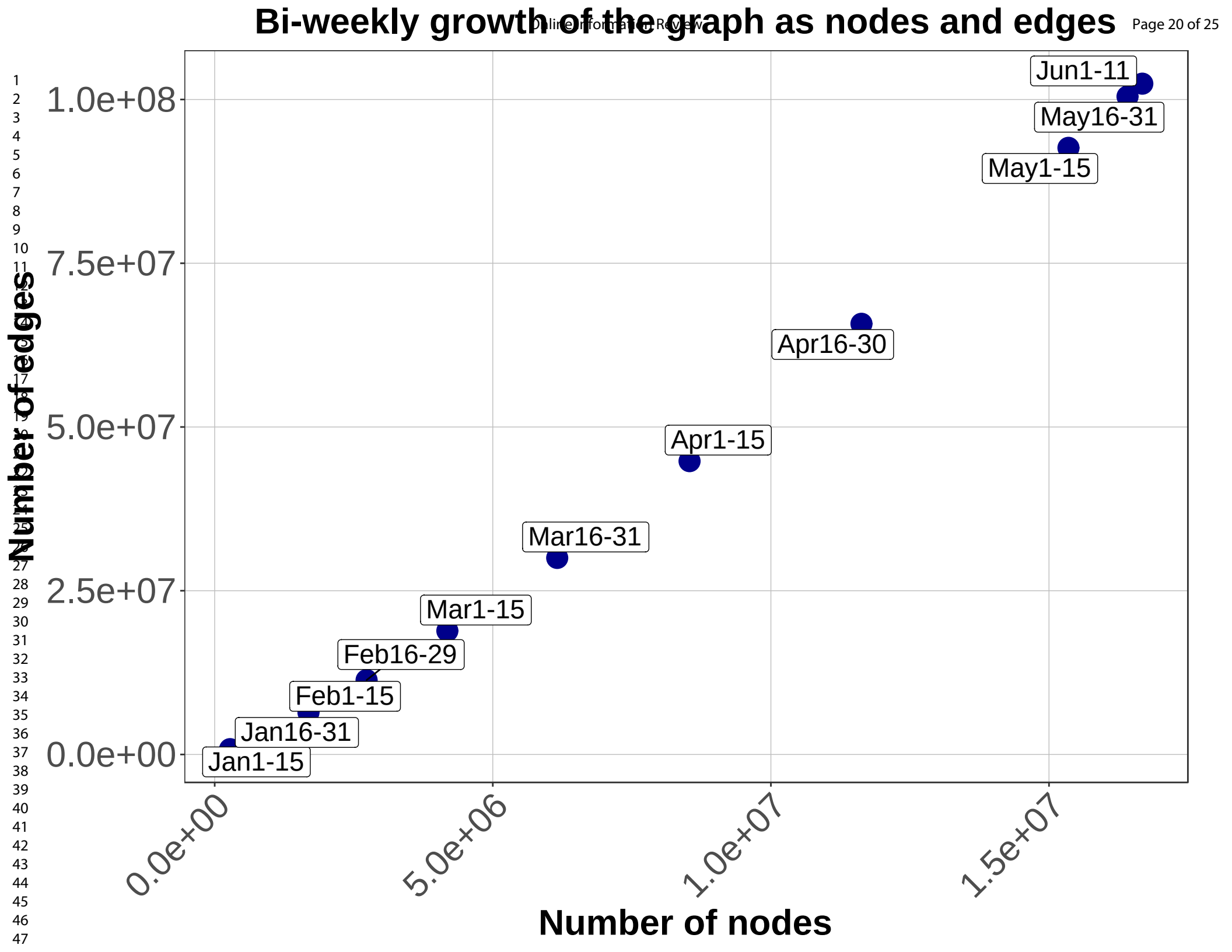
Each publication text is split into a list of sentences, which become the input for the following operations;
Initialize a list L_{res} to collect node pairs, therefore also the relations;
for *each sentence in the text* **do**
    Find the list of lemmatized biological entities present in the sentence with scispaCy = L_{bio};
    if $|L_{bio}| > 0$ **then**
        Find the list of prepositions present in the sentence with spaCy = L_{prep};
        for *each preposition in* L_{prep} **do**
            Find the head of the current parsing tree with spaCy = $left$;
            Find the first NOUN syntactic children of the current preposition with spaCy = $right$;
            if $left$ *OR* $right \in L_{bio}$ **then**
                Append ($left$, $right$) to the L_{res}
            end
        end
        Find the list of non-auxiliary verbs in the sentence with spaCy = L_{verbs};
        for *each verb in* L_{verbs} **do**
            Find the subject of the verb with spaCy = $subj$;
            Find the object of the verb with spaCy = obj;
            if $subj$ *OR* $obj \in L_{bio}$ **then**
                Append ($subj$, obj) to the L_{res};
            end
        end
    end
    Eliminate where either element of an (x, y) pair is empty or is a stopword from the L_{res};
    Eliminate cases where the elements of an (x, y) pair are the same (x ==y) from the L_{res};
end
Result: L_{res}

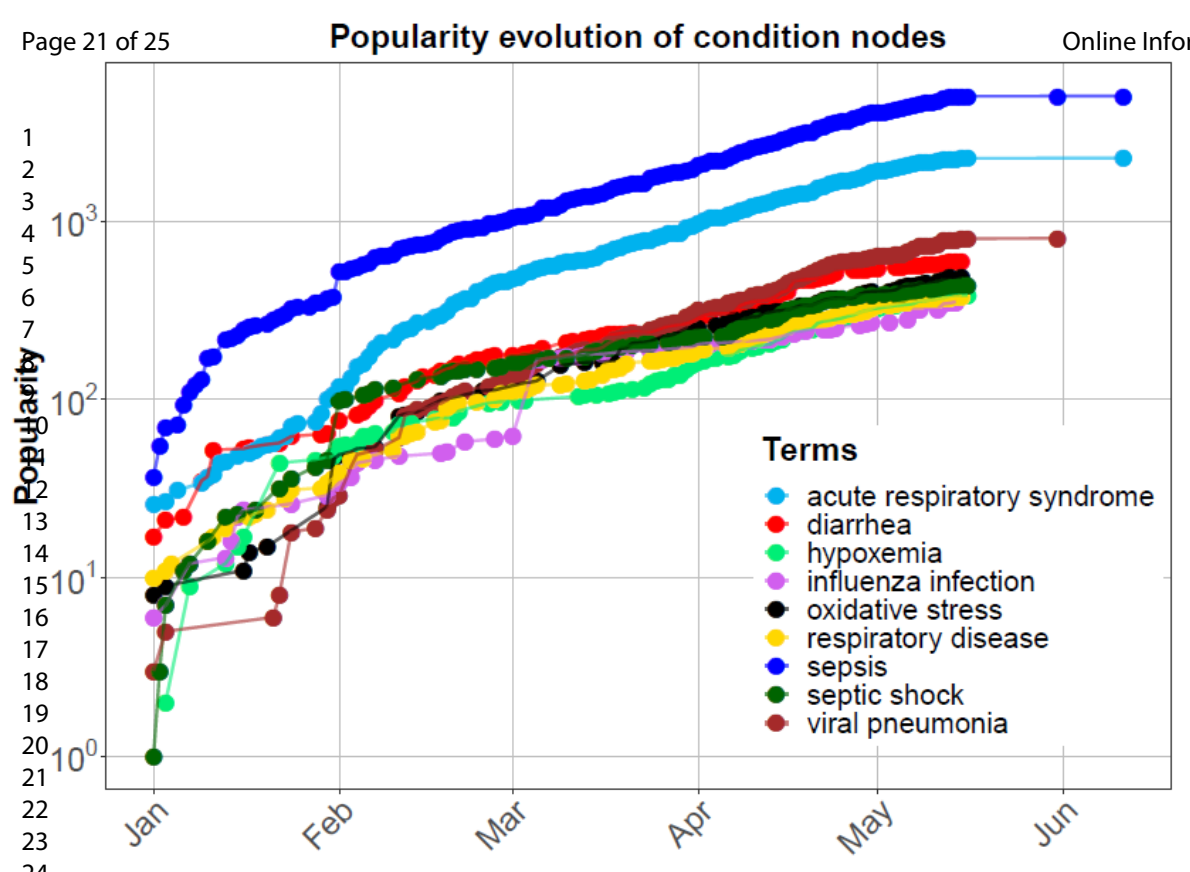
Online Information Review

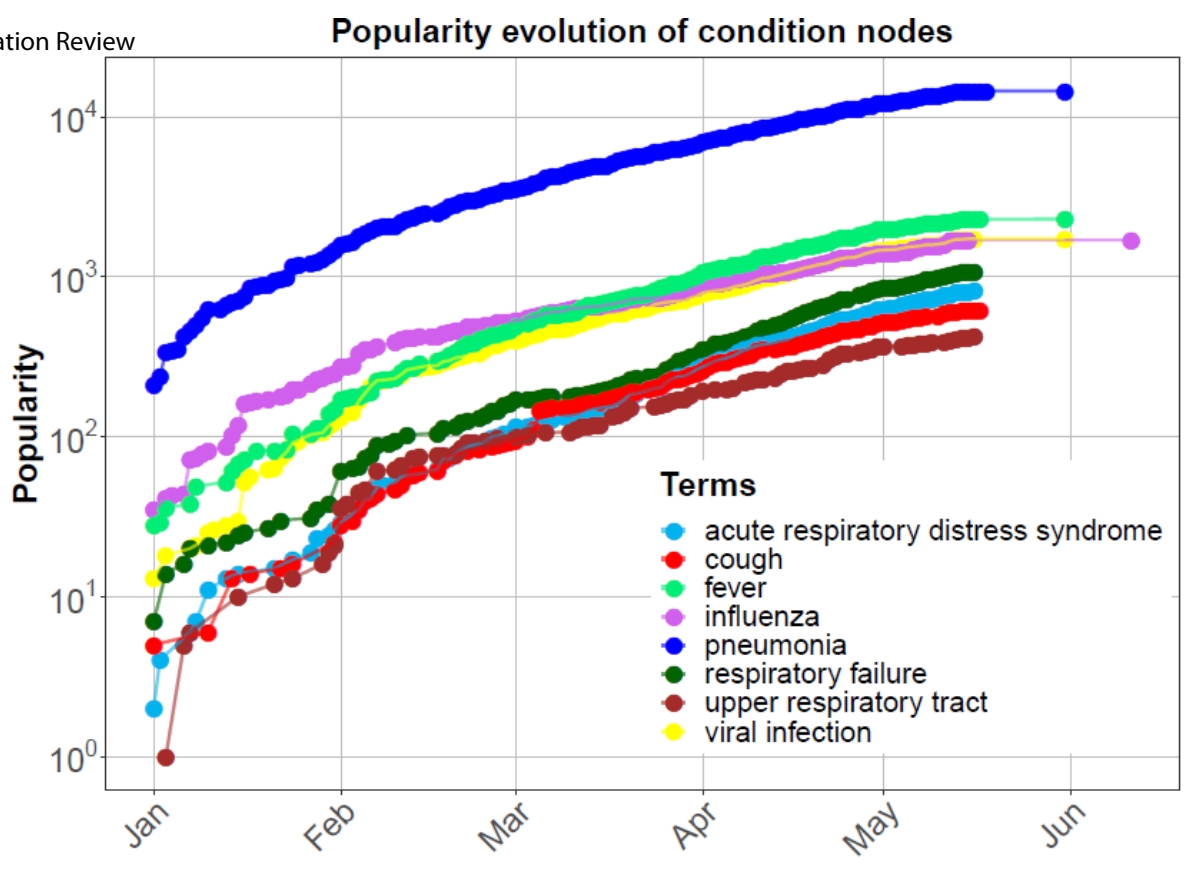
Publication dates of the studies used in the semantic graph

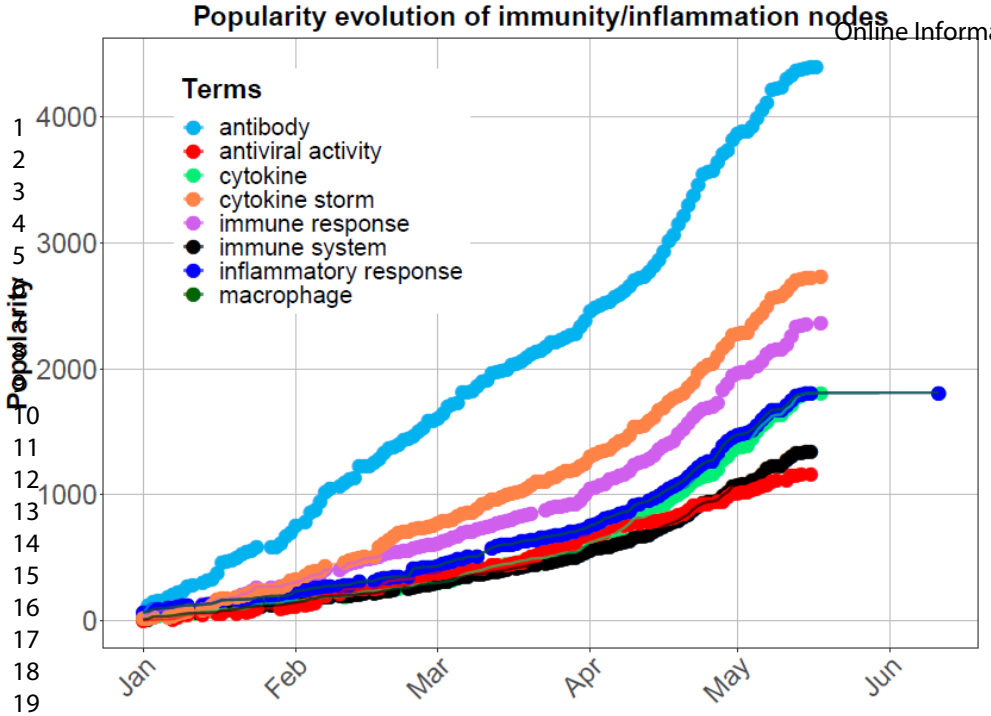


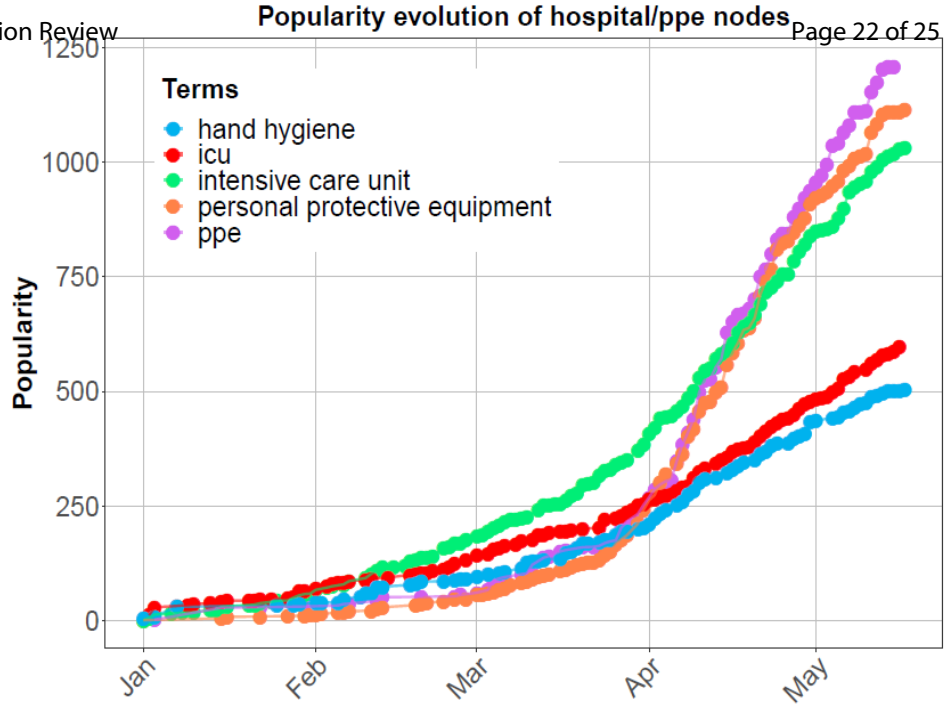

Evolution of average connectivity in the semantic graph
Average connectivity
7
6
5
4
3
Jan
Feb
Mar
Apr
May
Jun

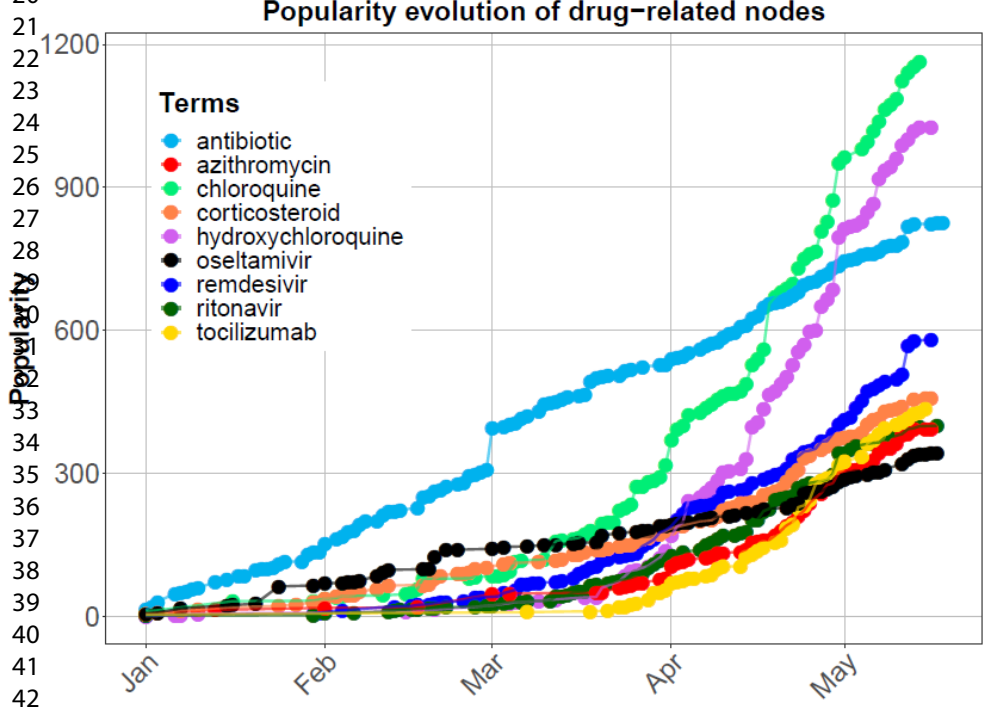

Bi-weekly growth of the graph as nodes and edges
Number of edges
Number of nodes
1.0e+08
7.5e+07
5.0e+07
2.5e+07
0.0e+00
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Jun1-11
May16-31
May1-15
Apr16-30
Apr1-15
Mar16-31
Mar1-15
Feb16-29
Feb1-15
Jan16-31
Jan1-15

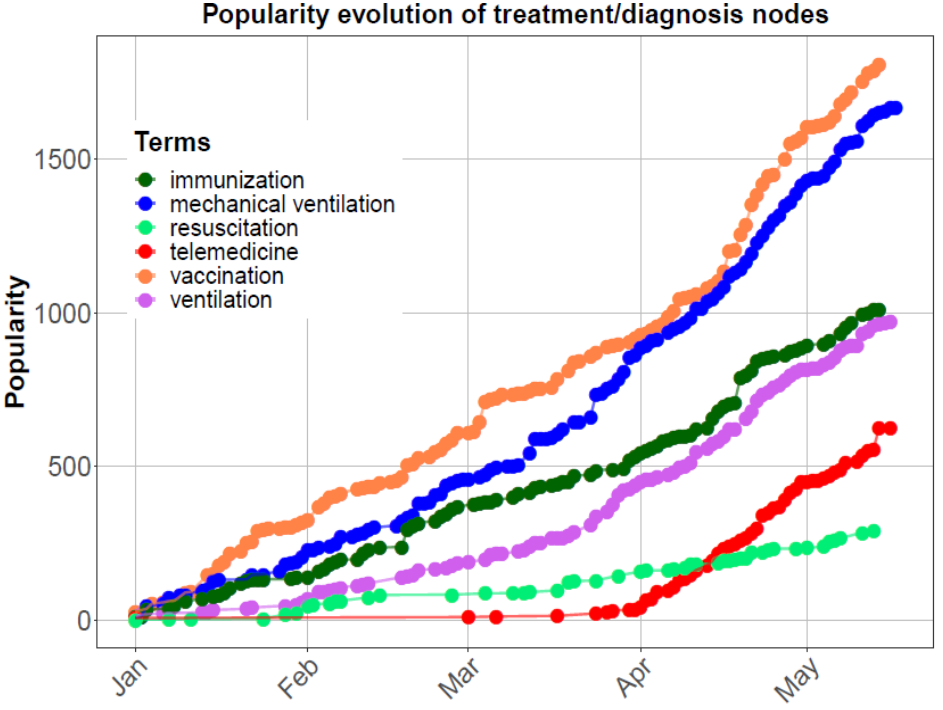

Popularity evolution of condition nodes
Popularity
10^3
10^2
10^1
10^0
Jan
Feb
Mar
Apr
May
Jun
Terms
acute respiratory syndrome
diarrhea
hypoxemia
influenza infection
oxidative stress
respiratory disease
sepsis
septic shock
viral pneumonia

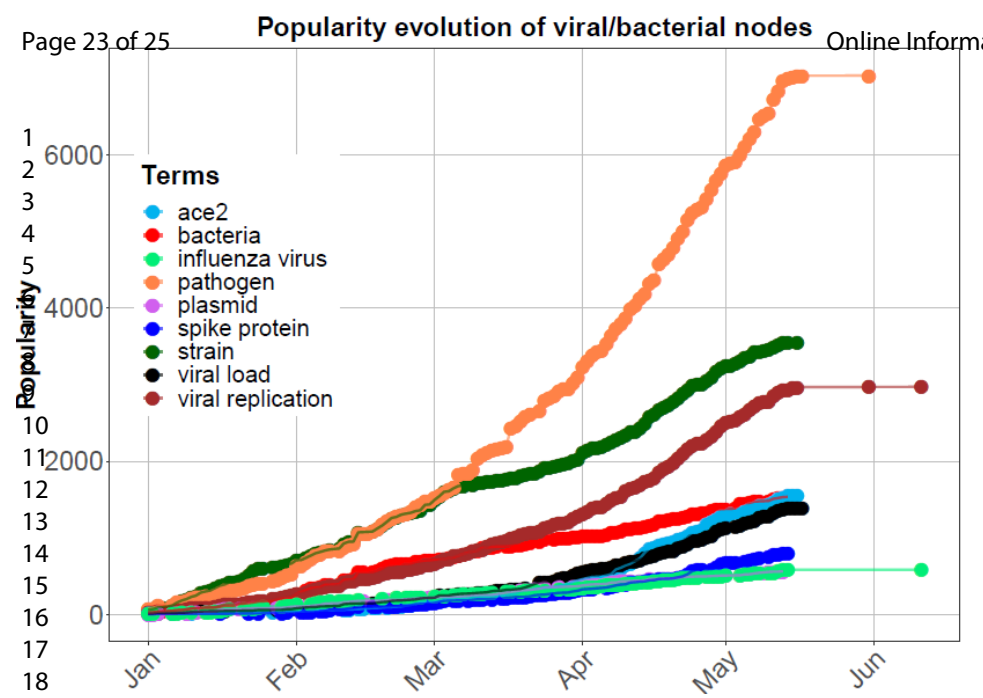

Popularity evolution of condition nodes
Popularity
10^4
10^3
10^2
10^1
10^0
Jan
Feb
Mar
Apr
May
Jun
Terms
acute respiratory distress syndrome
cough
fever
influenza
pneumonia
respiratory failure
upper respiratory tract
viral infection

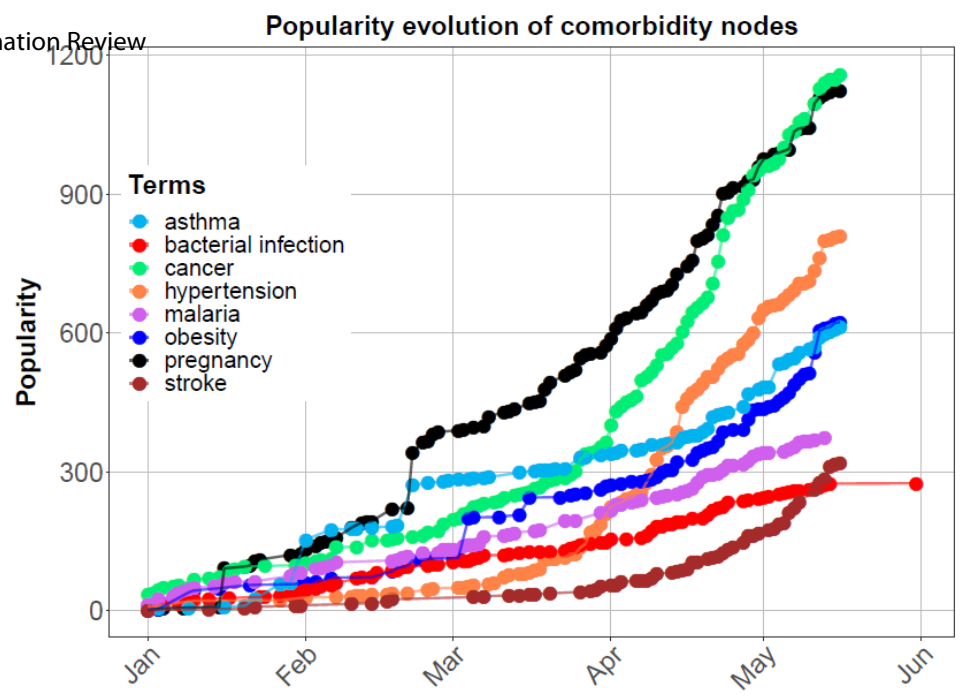

Popularity evolution of immunity/inflammation nodes
Terms
antibody
antiviral activity
cytokine
cytokine storm
immune response
immune system
inflammatory response
macrophage
Popularity
0
1000
2000
3000
4000
Jan
Feb
Mar
Apr
May
Jun

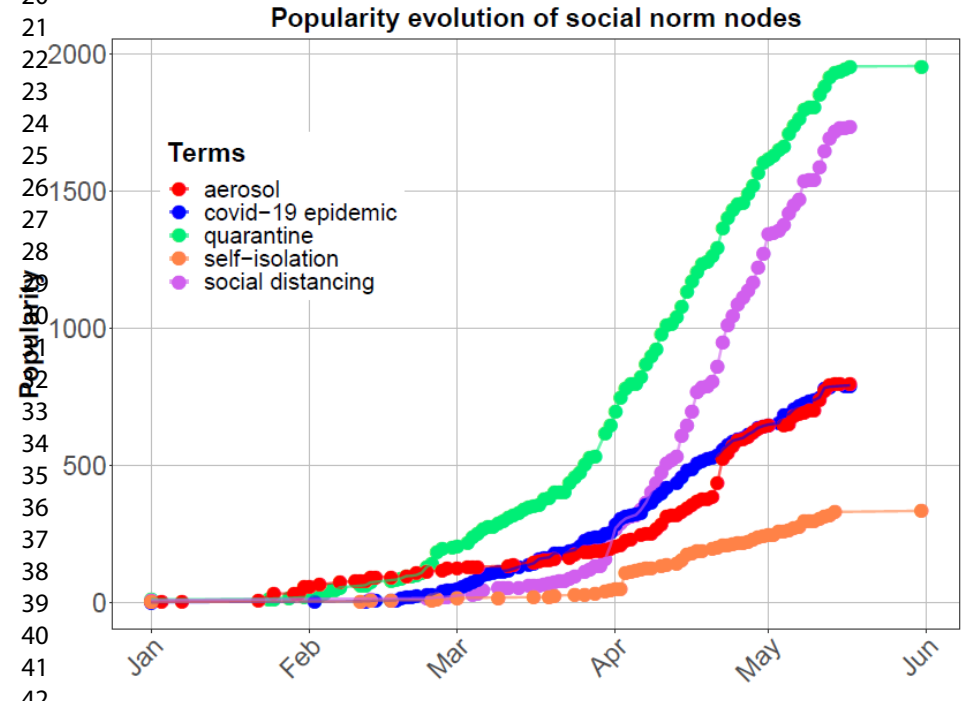

Popularity evolution of hospital/ppe nodes
Terms
hand hygiene
icu
intensive care unit
personal protective equipment
ppe
Popularity
0
250
500
750
1000
1250
Jan
Feb
Mar
Apr
May

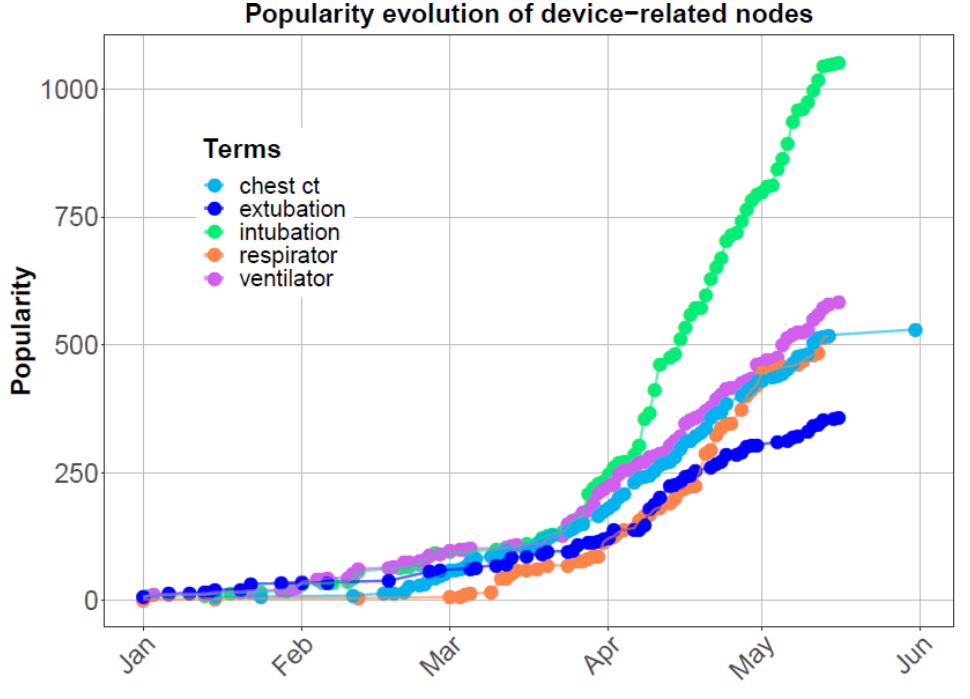

Popularity evolution of drug−related nodes
Terms
antibiotic
azithromycin
chloroquine
corticosteroid
hydroxychloroquine
oseltamivir
remdesivir
ritonavir
tocilizumab
Popularity
0
300
600
900
1200
Jan
Feb
Mar
Apr
May


Popularity evolution of treatment/diagnosis nodes
Terms
immunization
mechanical ventilation
resuscitation
telemedicine
vaccination
ventilation
Popularity
0
500
1000
1500
Jan
Feb
Mar
Apr
May


Popularity evolution of viral/bacterial nodes
Terms
ace2
bacteria
influenza virus
pathogen
plasmid
spike protein
strain
viral load
viral replication
Popularity
0
2000
4000
6000
Jan
Feb
Mar
Apr
May
Jun


Popularity evolution of comorbidity nodes
Terms
asthma
bacterial infection
cancer
hypertension
malaria
obesity
pregnancy
stroke
Popularity
0
300
600
900
1200
Jan
Feb
Mar
Apr
May
Jun


Popularity evolution of social norm nodes
Terms
aerosol
covid−19 epidemic
quarantine
self−isolation
social distancing
Popularity
0
500
1000
1500
2000
Jan
Feb
Mar
Apr
May
Jun


Popularity evolution of device−related nodes
Terms
chest ct
extubation
intubation
respirator
ventilator
Popularity
0
250
500
750
1000
Jan
Feb
Mar
Apr
May
Jun

Method		Classifier	
	Naive Bayes	Logistic Regression	Multi-layer Perceptron
Weighted Jaccard Coefficient	67.8 ± 0.10%	70.6 ± 0.32%	78.7 ± 0.5%
Weighted Adamic/Adar Index	79.1 ± 0.04%	74.2 ± 0.01%	78.7 ± 4.2%
Node2vec	69.6 ± 3.6%	73.7 ± 0.11%	79.0 ± 2.1%

Online Information Review

Entity-1	Entity-2	Link prediction score
sars-cov binding	nonsmall cell lung cancer	0.91
pbst buffer	reaction buffer	0.88
asialo-rhuepo	disease resistance	0.88
exercise	quinone reductase-2	0.87
anti-viral capability	cap-like structure	0.86
syncytiotrophoblast	virus rdrp	0.85
nucleotide identity matrix	pwm2 m	0.85
neuronal injury	specialist diagnosis	0.85
effector t cell activation	rituxan	0.85

Online Information Review