Network Analysis and Clustering of COVID-19 Literature

Zhiyi Sun, Yixuan Zeng

1 Introduction

In response to the pressing need for current and comprehensive insights on COVID-19 due to the overwhelming influx of new literature, this report embarks on a comprehensive network analysis of the COVID-19 literature data, utilizing Natural Language Processing (NLP) and AI and integrating Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and k-means clustering. PageRank algorithm and Latent Dirichlet Allocation (LDA) are considered for potential network analysis. The study distill complex patterns and relationships in COVID-19 literature data, offering visual and thematic insights into its research landscape. The analysis further aims to demonstrate the effectiveness of network analysis in organizing scientific data, aiding medical professionals and researchers in their pursuit of knowledge and solutions to COVID-19.

2 Methods

2.1 Data Prepossessing

The dataset used in this study is the COVID-19 Open Research Dataset Challenge (CORD-19) from Kaggle [`https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/data`]. The NLP technique is considered to extract text from documents, focusing on English language texts due to translation limitations. We clean this data by removing stopwords and use the encorescilg model from SpaCy for text analysis. The data is then converted to a numerical format using TF-IDF, facilitating text-based clustering. Additionally, we construct a biomedical knowledge graph for COVID-19, utilizing BioSNAP's collection of entities and identifying relationships among chemicals, genes, and diseases from scientific articles to create a network that captures complex biochemical interactions.

2.2 NetworkX and PageRank

NetworkX is a Python library acclaimed for its comprehensive toolset that enables analysis and manipulation of complex networks across diverse scientific fields. It facilitates the mapping of relationships and analysis of the dynamic interactions within data-rich structures. A notable feature of NetworkX is its implementation of the PageRank algorithm, which is mathematically represented by the formula[1]:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{1}$$

where $PR(p_i)$ denotes the PageRank of page p_i, d is the damping factor typically set to 0.85, N is the total number of pages, $M(p_i)$ is the set of pages linking to p_i, $PR(p_j)$ is the PageRank of page p_j, and $L(p_j)$ is the number of outbound links on page p_j.

PageRank is crucial in network analysis as it quantifies the significance of each node within a network based on its connections. This is particularly instrumental in biological network analysis where it aids in identifying key genes or proteins that may play central roles in the pathogenesis of diseases like COVID-19. NetworkX leverages this and other mathematical constructs to extract insights on network degrees, clustering, path lengths, and other topological features, which are indispensable for discerning the nuances of complex biological and biochemical networks.

2.3 Latent Dirichlet Allocation

LDA is a Bayesian network model commonly used in natural language processing and text mining to discover abstract topics within a collection of documents. In the context of analyzing COVID-19 literature, LDA can be instrumental in uncovering hidden thematic structures in large sets of scientific articles. By assuming that documents are mixtures of topics and topics are mixtures of words, LDA helps in categorizing text in a document into a certain topic based on word frequencies. The fundamental mathematics behind LDA involve probability distributions, particularly the Dirichlet distribution for topic assignment and multinomial distribution for word generation in topics[2]:

$$p(w|\theta, \beta) = \sum_{k=1}^{K} \theta_k \beta_{k,w} \tag{2}$$

where $p(w|\theta, \beta)$ denotes the probability of word w in a document. θ is the distribution of topics in the document. β is the distribution of words in topics. K is the number of topics. θ_k is the probability of topic k in the document. $\beta_{k,w}$ is the probability of word w in topic k.

2.4 Dimension Reduction, Clustering and Visualization

PCA is a statistical technique used to reduce the dimensionality of large datasets, while preserving as much variance as possible, by transforming a set of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component captures the most variance, followed by each successive component, under the constraint of being orthogonal to the previous ones. The core equation in PCA is the eigenvalue decomposition of the covariance matrix[3]:

$$\Sigma v = \lambda v \tag{3}$$

where Σ represents the covariance matrix, v is an eigenvector of Σ, and λ is the eigenvalue corresponding to the eigenvector v.

PCA is employed to effectively reduce the dimensionality of vectorized data X while retaining 95% of the variance, which assists in mitigating noise and outliers in the embedding Y_2. PCA serves as a preliminary step to simplify the complexity of our dataset and enhance the clarity of the subsequent clustering process. By selecting components that cumulatively account for 95% of the variance, we achieve a balance between dimensionality reduction and information retention.

Following PCA, the k-means clustering is applied to the reduced dataset to categorize the literature. The process involves assigning vectors to clusters based on their mean distance to iteratively updated centroids, with the number of clusters, k, being a critical parameter, which algorithm as shown in Figure 1. To determine the optimal value of k, we analyze the distortion - the sum of squared distances from each point to its assigned center - across different k values. The 'elbow' in the distortion plot, where further increases in k result in minimal decreases in distortion, indicates the most appropriate number of clusters.

```
Algorithm 1 K-Means
Require: Data matrix $A$ with rows $a_i$, $k$ the number of clusters
Randomly select $k$ data points from $A$ to start as cluster centroids $c_1, \dots, c_k$
do
    for $a_i \in A$ do                                          ▷ Assign data to clusters
        assign $a_i$ to $\pi_j$, where $c_j$ is the closest centroid to $a_i$
    end for
    for $j \in 1, \dots, k$ do                                  ▷ Update cluster centroids
        $\mu_j$ = the average of all $a_i \in \pi_j$
        $c_j = a_i$ such that $a_i$ is closest to $\mu_j$
    end for
while any $c_j$ updated or $a_i$ was assigned to a new cluster in the last iteration
return $\pi_i$ and $c_i$
```

Figure 1: K-means Algorithm

Subsequently, t-SNE is utilized for further dimensionality reduction, compressing the high-dimensional feature vector into 2 dimensions, thereby facilitating the visualization of the data in Figure 5. t-SNE aims to maintain the integrity of the high-dimensional relationships within a 2D space, allowing articles with similar topics to be positioned in closer proximity. Suppose there are n high-dimensional points, $x_1, x_2, ..., x_n$ and for a specific point i, point i picks point $j(\neq i)$ to be a neighbor with probability:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)} \tag{4}$$

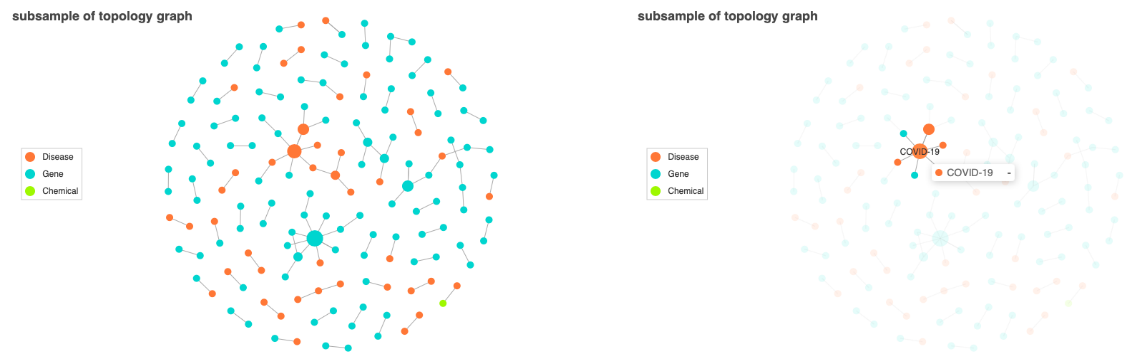
where σ_i(depends on i) controls the probability in which point j would be picked by i as a neighbor. Consequently, the clusters found by k-means were used as labels to help visually separate different concentrations of topics (Figure 2).

```
Algorithm 1: Clustering and Visualization
Input: Data matrix $\mathbf{X}_{m\times n} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^{m\times n}$ where $m$ rows are genes and $n$ columns are cells.
Output: Cell clusters and a low dimensional projection
Compute the sample mean $\mu_n$ and the centered matrix $\mathbf{X}_c = \mathbf{X} - \mu_n \mathbf{1}^\top$ where $\mathbf{1}$ is a vector of ones
Compute the SVD of $\mathbf{X}_c = \mathbf{U\Sigma V}^\top$
Construct $\mathbf{P}_{n\times r} = \begin{bmatrix} v_1 & v_2 & \dots & v_r \end{bmatrix}$ where each column in $\mathbf{P}$ is a right singular vector of $\mathbf{X}_c$. Here $r$
  can be chosen using the optimal hard threshold [3] on $\mathbf{X}_c$
Construct a similarity matrix $\mathbf{A}_{n\times n}$ from $\mathbf{P}$ by determining the distance between each row. The choice
  of distance measure depends on the data type and user preference. Examples include Gaussian
  similarity, Euclidean distance, Manhattan distance (city block distance), Kullbeck-Liebler divergence,
  and correlation
Perform clustering: spectral or modularity clustering on $\mathbf{A}$ with $k$ clusters. $k$ can be chosen using
  domain knowledge or by testing multiple values of $k$ and evaluating the best performance. Note: $k$
  may be $\leq r$
Visualization: t-SNE or UMAP to reduce the dimensions of $\mathbf{P}$ and visualize data colored according to
  clusters
```

Figure 2: Clustering and Visualization

3 Results

To analyze the data, we employed several computational methods to analyze a biomedical knowledge graph related to COVID-19, aiming to understand the intricate web of associations between chemicals, genes, and diseases. The transformed data, comprising 40,812 rows and 5 columns (start entity, end entity, start entity type, end entity type, and marked sentence), encapsulated a network with 3,644 unique nodes and 20,484 edges (Figure 3).



(a) Subsample of Topology Graph (Disease, Gene & Chemical)

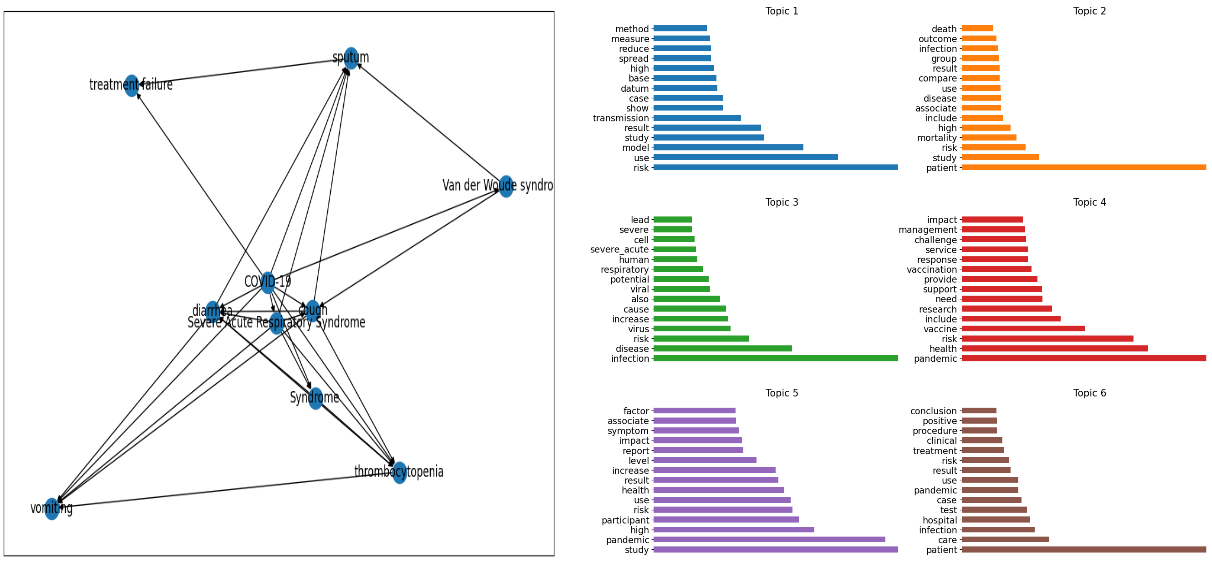
(b) Subsample of Topology Graph (Covid-19)

Figure 3: Topology Graph

This network analysis showed it's not strongly connected, meaning not all nodes are interconnected. The largest strongly connected component had an average shortest path length of 0, indicating direct connections, which may suggest direct relationships between certain COVID-19 symptoms or genes. With an average degree of 11.2426, the network displayed a moderate level of interaction, possibly reflecting the complex interaction between genetic factors and symptoms in COVID-19. Transitivity and average clustering coefficients of 0.1267 and 0.2250, respectively, indicated moderate clustering, mirroring how certain symptoms or genetic factors are associated with specific COVID-19 manifestations. The low network density of 0.0015 suggests sparse connections, typical of biological systems with many unrealized potential interactions.

Using the PageRank algorithm, we found each node's importance based on connections, with a degree assortativity of -0.071, typical for biological networks where high-degree hubs connect to low-degree non-hubs. In COVID-19, this implies certain symptoms or genes are central in disease progression. The top 10 nodes - 'vomiting', 'sputum', 'Severe Acute Respiratory Syndrome', 'diarrhea', 'thrombocytopenia', 'treatment failure', 'COVID-19', 'cough', 'Syndrome', and 'Van der Woude syndrome' - highlight key clinical and pathological aspects of the disease, suggesting areas for focused medical research and drug development. The less influential nodes, including various genes and chemicals, might represent more peripheral elements of COVID-19 or lesser-understood aspects in current research. In terms of the structural robustness and influence distribution, the investigation revealed that a few nodes hold a disproportionate amount of influence, as indicated by the distribution of PageRank values. Efforts to identify removable edges without causing network fragmentation were unsuccessful, indicating no expendable interactions and underscoring the network's tight interconnectivity. This characteristic, along with the observed negative degree assortativity common in biological systems, highlights the importance of central nodes in maintaining network integrity.

LDA analysis streamlined COVID-19 literature into six key topics with a high coherence score, underscoring comprehensive research on treatments like vaccines and drugs, the psychological impact under 'mental health', and the clinical gravity shown by 'severe' and 'infection'. The literature also addresses the pandemic's transmission and the societal response in public services (i.e., 'care' and 'service'), with 'mortality' and 'death' terms highlighting the ultimate toll. This thematic distribution across articles illustrates a well-rounded scientific investigation into the pandemic's multifarious effects (Figure 4).

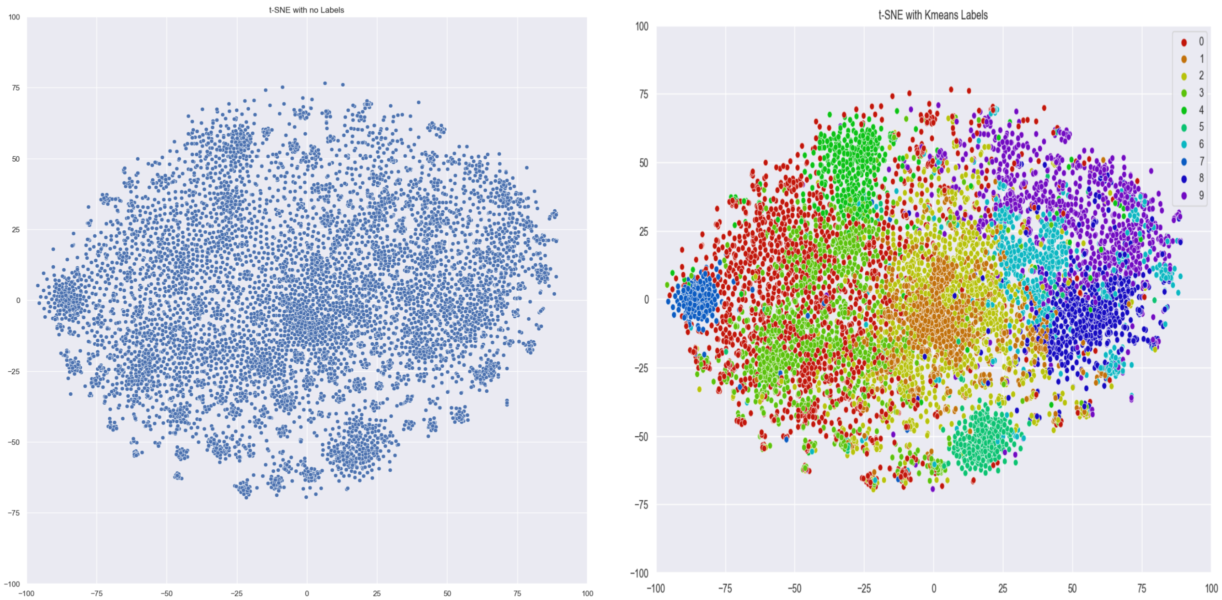


(a) Top 10 PageRank Diseases

(b) LDA: 6 Topics

Figure 4: PageRank Diseases and LDA

For clustering and visualization, the Figure 5 (a) created by t-SNE reveals a dense core of closely related documents and scattered outliers, indicating varied document similarity. Without labels, it's difficult to draw specific conclusions about the nature of the clusters or the relationships between different documents.



(a) t-SNE without Labels

(b) t-SNE with Kmeans Labels

Figure 5: t-SNE Visualization

Therefore, to add more interpretive value, we apply k-means to label these points. The labeled Figure 5 (b) gives better insight into how the papers are grouped. Interestingly, both k-means and t-SNE can find independent clusters even though they were run independently. This shows that structure within the literature can be observed and measured to some extent. It is worth mentioning that the spread of color-coded labels across the plot reflects the intricate connections and thematic overlaps found in the higher-dimensional space, which is a result of t-SNE and k-means finding different connections in the higher dimensional data. The topics of these papers often intersect so it was hard to cleanly separate them.

4 Conclusion

The study's network analysis of COVID-19 literature revealed a moderately connected landscape, featuring a mix of chemicals, genes, and diseases. This highlights a broad, yet not uniformly interconnected research area. Notably, a strongly connected component with an average shortest path length of zero indicates direct relationships, crucial for understanding COVID-19's genetic and symptomatic links. Despite the network's low density, moderate clustering was observed, suggesting complex interactions within the disease's pathophysiology. The PageRank algorithm identified important nodes like 'vomiting', 'cough', and 'COVID-19', which are vital in understanding the disease's progression and potential treatments, while less influential nodes may reveal lesser-known aspects of the disease. LDA analysis efficiently condensed the literature into six key topics, including clinical treatments, psychological impacts, and transmission dynamics, reflecting the pandemic's multifaceted nature and the depth of scientific exploration.

In addition to network analysis, the study employed NLP and machine learning to further refine the COVID-19 literature into a more structured and coherent network. PCA was instrumental in reducing data complexity, while k-means clustering helped organize the data, revealing key themes within the pandemic's discourse. The integration of t-SNE with k-means clustering provided a more nuanced view of thematic groupings, although the complexity of the data sometimes made categorizations challenging. This methodological approach not only confirmed the relevance of the clustered articles but also highlighted the effectiveness of this analytical technique in network analysis. These insights suggest that interventions targeting pivotal nodes could be key in disrupting COVID-19's progression. The study offers valuable guidance for health professionals and researchers, with future enhancements potentially including the application of cosine similarity in graph theory and the development of user-friendly interfaces for navigating the dataset network.

5 Reference

[1] Boldi, P., Santini, M., & Vigna, S. (2005, May). PageRank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web* (pp. 557-566).

[2] Teh, Y., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in neural information processing systems*, 19.

[3] Huang, L., Nguyen, X., Garofalakis, M., Jordan, M., Joseph, A., & Taft, N. (2006). In-network PCA and anomaly detection. *Advances in neural information processing systems*, 19.

[4] Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, 307, 39-52.