Real-time and Explainable Detection of Epidemics with Global News Data

Sungnyun Kim [*1] **Jaewoo Shin** [*1] **Seongha Eom** [1] **Jihwan Oh** [1] **Se-Young Yun** [1]

Abstract

Monitoring and detecting epidemics are essential for protecting humanity from extreme harm. However, it must be done in real time for accurate epidemic detection to use limited resources efficiently and save time preventing the spread. Nevertheless, previous studies have focused on predicting the number of confirmed cases after the disease has already spread or when the relevant data are provided. Moreover, it is difficult to give the reason for predictions made using existing methods. In this study, we investigated how to detect and alert infectious diseases that might develop into pandemics soon, even before the information about a specific disease is aggregated. We propose an explainable method to detect an epidemic. This method uses only global news data, which are easily accessible in real time. Hence, we convert the news data to a graph form and cluster the news themes to curate and extract relevant information. The experiments on previous epidemics, including COVID-19, show that our approach allows the explainable real-time prediction of an epidemic disease and guides decision-making for prevention. Code is available at https://github.com/sungnyun/Epidemics-Detection-GKG.

1. Introduction

Since COVID-19 (SARS-CoV-2, or Severe Acute Respiratory Syndrome Coronavirus 2) began to spread in early 2020, the disease has harmed people worldwide. It has disrupted the lives and health of many people. By the end of 2021, more than 288 million people had been infected, and 5.4 million had died (Ritchie et al., 2020). Even today, in 2022, COVID-19 is not disappearing, as various mutations are arising, and some countries (*e.g.*, South Korea, Australia, and Thailand) are still not free from COVID-19 policies. The disease restricted the movement of individuals between countries, and people were required to spend a significant amount of time in quarantine. The implementation of social distancing, self-isolation, and movement restrictions reduced the economic output of workers and companies, resulting in the collapse of the global supply chain and a decrease in global trade (Ozili & Arun, 2020). As a result, COVID-19 continues to have global adverse effects, such as social anxiety and a depressed economic growth rate.

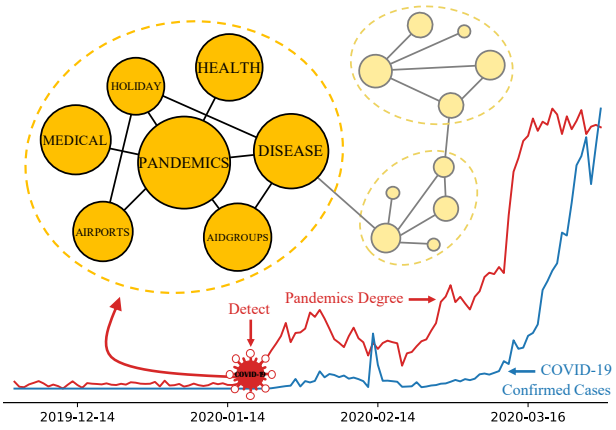


Figure 1. We can detect infectious diseases in real time using the pandemics degree (Section 5.1) and explain the detection by theme graph clustering to identify the alarming factors (Section 5.2). Moreover, extracting valuable information about the disease is possible even before its confirmed cases are counted (Section 5.3).

From the end of 2019 when the disease first broke out, there were a few weeks of a response period before World Health Organization (WHO) declared it a pandemic on March 11, 2020. However, most countries were late to recognize it, and this late response resulted in national losses (Bosa et al., 2022). If it had been possible to detect in advance that the diseases might be declared a pandemic, the damage could have been reduced. In this sense, early detection of infectious diseases is an important study because we can save time preparing before the spread and use limited medical resources more efficiently. Nevertheless, existing studies focus on developing models to predict the number of infected people *after* it has spread rather than detecting the epidemic *before* it spreads worldwide. Moreover, existing studies on the early detection of epidemics (Hashimoto et al., 2000; Feng et al., 2021) need a high level of domain knowledge.

[*]Equal contribution [1]Kim Jaechul Graduate School of Artificial Intelligence, KAIST, Seoul, South Korea. Correspondence to: Se-Young Yun <yunseyoung@kaist.ac.kr>.

Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022.

Therefore, we explore how to detect infectious diseases that can develop into pandemics without prior information on a specific disease or high-level domain knowledge. Lucas et al. (2020) and Aiken et al. (2020) insisted that keyword search volume, one of the real-time online data, is highly correlated with the number of infected cases. However, dataset bias limits the flexible use of search data, such as that search keywords differ from person to person, and English keywords have high volumes only in English-speaking countries. Instead, we use open-source global news data to capture information from worldwide real-time reports. We confirmed in this study that early detection is possible through a simple graph clustering of the news data. Specifically, it can serve as a more explainable model by extracting the themes related to infectious diseases.

Our first contribution is that we demonstrate that simply measuring the frequency of pandemic-related news can detect the spread of the disease; however, the enormous amount of worldwide news makes it impossible for a human to detect every news item that contains important information. Moreover, if the result cannot be explained, we cannot provide the reason why the news frequency rises and thus provide a platform leading to prevention policies. Therefore, our additional contributions are that we perform explainable epidemic detection by converting the news data to graph form and clustering the news themes to curate the relevant information. Figure 1 provides an overview of our process, which depicts detecting the COVID-19 epidemic as well as finding an informative cluster to help us respond to the disease.

The remainder of the paper is organized as follows. Section 2 reviews previous studies related to the subject. Section 3 outlines our data-gathering and -preparation methodology. In Section 4, we demonstrate that we can predict the spread of the disease using the amount of pandemic-related news. In Section 5, we propose an explainable method to detect epidemics and analyze the results based on graph clustering. By conducting various case studies on diseases, we figure out which information we could extract and use for policy-making. Finally, in Section 6, we confirm that the extracted themes improve the forecasting of COVID-19 cases.

2. Related Works

2.1. Epidemic Modeling and COVID-19 Forecasting

While epidemic modeling has long been studied as a research subject, mainly with compartmental and machine learning models, the current interest in infectious diseases has triggered more advanced methodologies. Many epidemic-modeling works on COVID-19 have attempted to model the dynamics of the disease with a compartmental model (*e.g.*, SIR) and its variants (*e.g.*, SEIR and SEIRD[1]) (Hao et al., 2020; Yang et al., 2020; Du et al., 2021; Korolev, 2021). Several of these works extend the epidemic modeling in combination with machine learning or deep learning methods to predict the infected population better. They use a regression model (Korolev, 2021), a probabilistic graphical model (Qian et al., 2020; Vega et al., 2022), or a neural network (Deng et al., 2020; Arik et al., 2020; Menda et al., 2021), to estimate the essential parameters in the SIR variants. It is easy to interpret the dynamics of each compartment in such algorithms. However, they have critical issues in that they require numerous hypotheses based on domain knowledge, and the results vary significantly depending on the design choice of the simulators and hypotheses (Korolev, 2021; Abbasimehr & Paki, 2021).

Thus, more studies have focused recently on developing fully data-driven approaches, mostly with deep learning models. Chimmula & Zhang (2020) attempted to overcome the limitations of statistical models by using a Long Short-Term Memory model to use real-time data. Kim et al. (2020) used Transformer with flight data to predict the confirmed cases of overseas inflow. Because disease propagation patterns are similar across different regions in different time periods, Jin et al. (2021) applied cross-attention to encode inter-regional dependencies.

The recent trend is moving from compartmental models to deep learning models with sufficient well-refined data to predict outcomes more accurately. However, previous works have focused on predicting confirmed cases, which requires statistical data (*e.g.*, previous confirmed cases) or background knowledge (*e.g.*, clinical diagnosis of the patients), thus inhibiting the real-time detection of epidemics before the spread.

2.2. Research Scope with News Dataset

News article datasets have often been used for prediction tasks. Jacobs et al. (2018) proposed a strategy for predicting economic events from English news articles using supervised classification. In particular, several works used the Global Database of Events, Language, and Tone (GDELT) dataset for prediction. Galla & Burke (2018) discovered and predicted social unrest at the county level to deploy programs and applications to mitigate its negative consequences, while Jakel (2019) predicted stock prices on a selection of companies.

Since the COVID-19 crisis, studies have been proposed to use news datasets for various tasks regarding COVID-19. They have extracted various features from news articles that include rich contextual information (*e.g.*, tone, entity,

[1]Each compartment corresponds to Susceptible, Exposed, Infected, Recovered, and Deceased.

Table 1. Examples of integrating subthemes into superthemes based on our rule.

Supertheme	←	Subthemes
AIDGROUPS	←	TAX-AIDGROUPS · TAX-AIDGROUPS-RED-CROSS · TAX-AIDGROUPS-UNICEF · TAX-AIDGROUPS-WORLD-HEALTH-ORGANIZATION · · ·
DISEASE	←	TAX-DISEASE-DISEASE · TAX-DISEASE-CANCER · TAX-DISEASE-INFECTION · TAX-DISEASE-OUTBREAK · TAX-DISEASE-EMERGENCIES · TAX-DISEASE-FEVER · TAX-DISEASE-FLU · TAX-DISEASE-BACTERIA · TAX-DISEASE-TRAUMA · TAX-DISEASE-CORONAVIRUS · · ·
FNCACT	←	TAX-FNCACT · TAX-FNCACT-PRESIDENT · TAX-FNCACT-CHILDREN · TAX-FNCACT-FACULTY · TAX-FNCACT-MANAGERS · TAX-FNCACT-LEADER · TAX-FNCACT-CRIMINAL · TAX-FNCACT-BUSINESS-EXECUTIVES · TAX-FNCACT-MEDICAL-WORKERS · · ·
HEALTH	←	GENERAL-HEALTH · WB-1287-HEALTH-INSURANCE · WB-1331-HEALTH-TECHNOLOGIES · WB-2165-HEALTH-EMERGENCIES · · ·
PANDEMICS	←	WB-2167-PANDEMICS
QUARANTINE	←	SOC-QUARANTINE
TRANSPORT	←	PUBLIC-TRANSPORT · WB-135-TRANSPORT · WB-793-TRANSPORT-AND-LOGISTICS-SERVICES · WB-1803-TRANSPORT-INFRASTRUCTURE · · ·

and event tags). Krawczyk et al. (2021) quantified the total volume of social media, government pages, and COVID-19 articles and discovered that news articles provided the most reliable information. Some studies have used the GDELT dataset. For example, Shahsavari et al. (2020) used it to show how conspiracy theories about COVID-19 rely on disparate knowledge domains and discuss how they relate to broader pandemic reporting. Fu & Zhu (2020) used the GDELT dataset to evaluate the Chinese media transparency. Chakraborty & Bose (2020) discovered a high correlation between news sentiment and COVID-19 statistics. However, even though GDELT contains useful signals for predicting social phenomena, none of these studies used the GDELT news dataset to predict or detect epidemic outbreaks.

3. Data Preparation and Processing

In Sections 3.1 and 3.2, we describe how the news data were collected and preprocessed, and in Section 3.3, we introduce the clustering algorithm we used.

3.1. Datasets and Graph Generation

We used the Global Knowledge Graph (GKG) dataset from *The GDELT Project*[2] for our experiments. The GKG dataset gathers worldwide news data for each date, incorporating every person, organization, theme, location, tone, news source, and event across the planet into a single massive set that captures what is happening around the world, its context, the people involved, and the way people worldwide feel about it. We use three variables: *Themes*, *Locations*, and *Tone*. Relevant themes and locations are tagged for each news article, and the tone values are recorded as floating-point numbers representing the news emotion in six dimensions. For more information about the GKG dataset, refer to Saz-Carranza et al. (2020). For evaluation, we also needed confirmed cases of diseases. For COVID-19, we used the OWID (Our World in Data) dataset (Ritchie et al., 2020), and for the Ebola virus, we used monthly cases data provided by the CDC (Centers for Disease Control and Prevention).[3]

[2] https://www.gdeltproject.org

[3] https://www.cdc.gov

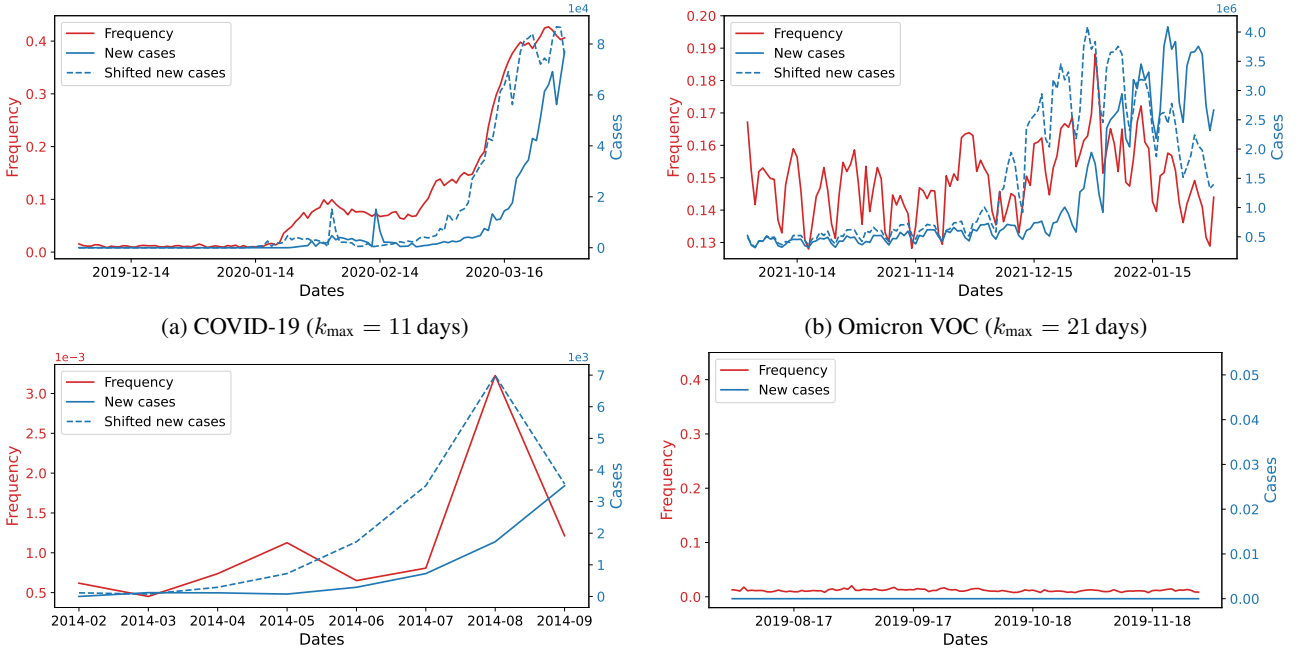
Table 2. Time and memory cost by graph type. **Type E**: Using the entire subthemes (*i.e.*, the original graph). **Type R**: Reducing to superthemes. **Type D**: Disassembling only the DISEASE supertheme (Section 5.3). The number of edges, graph generation time, and graph clustering time are different for each day, so we averaged seven daily graphs (*i.e.*, one week): 02/01/2020–02/07/2020.

Type	E	R	D
# Nodes	56,840	1,125	5,549
# Edges	2.44 M	0.18 M	0.28 M
Graph generation time (s)	501.8	27.1	27.6
Graph clustering time (s)	105.1	3.2	7.2
Graph size (MB)	24,649	10	235

We convert the GKG news data into a graph form. We consider a daily hypergraph $G_H(t) = (\mathcal{V}, \mathcal{E}_H(t))$ for day t (or step t), where each node $v \in \mathcal{V}$ represents a theme and a hyperedge $e \in \mathcal{E}_H(t)$ represents a daily news article containing a subset of themes. However, clustering with hypergraphs, especially with real-world, large-scale graphs, has not been sufficiently studied (Chodrow et al., 2021). Instead, we use a clique expansion (Zhou et al., 2006; Benson et al., 2016) and consider a weighted graph $G(t) = (\mathcal{V}, \mathcal{E}(t))$ with $|\mathcal{V}| = n, |\mathcal{E}(t)| = m_t$ for each day t. The edge set is constructed by replacing each hyperedge with a clique, that is, $\mathcal{E}(t) = \{(i, j) \,|\, i, j \in e, e \in \mathcal{E}_H(t)\}$. The edge weight is proportional to the number of hyperedges (*i.e.*, news) that contain the two connected nodes (*i.e.*, themes), and the weights are normalized by the maximum edge weight.

3.2. Theme Hierarchy

The GKG news dataset contains 56,840 themes ($n = 56,840$). The dataset names a theme by listing categories in descending order by the number of items. For example, TAX-AIDGROUPS-RED-CROSS is a *subtheme* included in the larger category TAX-AIDGROUPS. We name the category AIDGROUPS without TAX- as a *supertheme*. To reduce excessive computation, we integrate several subthemes into a supertheme. For example, TAX-DISEASE-CANCER is included in the supertheme DISEASE, while single-category themes, such as LEADER or ARREST, are superthemes themselves. Table 1 lists other examples of integrating subthemes. After integrating all subthemes according to the



(a) COVID-19 ($k_{\max} = 11$ days)

(b) Omicron VOC ($k_{\max} = 21$ days)

(c) Ebola virus ($k_{\max} = 2$ months)

(d) No epidemics

Figure 2. The news frequency measured over four different epidemiological periods. We considered four months for each period, except for the Ebola virus case, where only the monthly cases were available, and the disease spread slowly. The specific dates are as follows: (a) 12/01/2019–03/31/2020, (b) 10/01/2021–01/31/2022, (c) 02/01/2014–09/30/2014, and (d) 08/01/2019–11/30/2019. A dashed line indicates the graph of newly infected cases shifted by $k_{\max}$ days (or months for EBOV) of the pandemics frequency.

theme hierarchy, we came up with 1,125 superthemes. For the remainder of our paper, unless specified, a theme denotes a supertheme. We also provide the code[4] to reproduce our results reducing themes.

Reducing the nodes makes our graph generation and clustering much more efficient. Table 2 summarizes the decrease in computation and memory cost when using superthemes rather than all subthemes. When we use the entire subthemes (**Type E**), the cost is significant, and the higher number of edges makes it challenging to determine which links are important. Reducing the subthemes to superthemes (**Type R**) allows us to generate graphs quickly and analyze them, although it can have a limitation in that a detailed analysis becomes difficult due to the compression of information.

Because of the trade-off between the cost and information loss, we propose a different type of compression, where we disassemble only the DISEASE supertheme (**Type D**) from the **Type R** graph. The cost slightly increases, but it provides enough detail on the themes related to diseases. This strategy is discussed in depth in Section 5.3.

3.3. Louvain Algorithm (Blondel et al., 2008)

Clustering the news themes allows us to see which themes are most relevant in the graph-form data generated above. Therefore, we use the Louvain algorithm (Blondel et al., 2008) for our clustering method because it is a fast method for detecting communities in large weighted networks. The Louvain algorithm works by iteratively finding the communities with the maximum modularity, which quantifies how densely the nodes in the communities are connected:

$$Q = \frac{1}{2m} \sum_{i,j}^{n} \left(A_{i,j} - \frac{d_i d_j}{2m} \right) \cdot \sigma[C(i), C(j)], \quad (1)$$

where m and n denote the number of edges and nodes, respectively, and d_i is the degree of node i. $\mathbf{A}$ is an adjacency matrix in which $A_{i,j} \in \{0, 1\}$ has the value 1 if node i and j are connected, and $\sigma[C(i), C(j)] \in \{0, 1\}$ has the value 1 if node i and j are in the same community.

4. News Frequency for the Early Detection of Epidemics

We can use the news frequency to detect epidemics. News frequency is defined as the fraction of daily news with a particular theme. For example, if 10% of the day t news items contain the theme PANDEMICS (the corresponding subtheme is WB-2167-PANDEMICS), the pandemics frequency f_t is 0.1. Figure 2 shows that measuring the pandemics frequency, *i.e.*, simply counting the number of news items tagged with PANDEMICS, can detect the prevalence of disease in advance.

[4] https://github.com/sungnyun/Epidemics-Detection-GKG

Table 3. Evaluation of the pandemics frequency for the infected cases of each disease in Figure 2. Bold type corresponds to the maximum shift days and the maximum correlation.

Disease	Eval.	Shifting steps and correlation				
COVID-19	k	9	10	**11**	12	13
	ρ	96.38	96.76	**97.14**	96.56	96.65
Omicron VOC	k	19	20	**21**	22	23
	ρ	45.52	52.12	**56.66**	49.26	42.21
Ebola virus	k	0	1	**2**	3	4
	ρ	46.76	47.05	**86.73**	26.25	13.26

We propose two metrics to evaluate how accurate and timely this frequency is in detecting the epidemic: the maximum correlation ($\rho_{\max}$) and maximum shift days ($k_{\max}$). Formally, they are defined as:

$$\rho_{\max} = \max_{k \geq 0} \rho\big(f_{t:(t+L)}, c_{(t+k):(t+L+k)}\big) \tag{2}$$

$$k_{\max} = \arg\max_{k \geq 0} \rho\big(f_{t:(t+L)}, c_{(t+k):(t+L+k)}\big) \tag{3}$$

where f_t and c_t are the daily frequency and daily confirmed cases, respectively. L is the evaluation sequence length, and k is the shift in days (or steps). A high $\rho_{\max}$ value implies that the frequency is well matched to $k_{\max}$-days future cases.

We measured the news frequency in four different periods. Figure 2a depicts the outbreak of COVID-19. Before the disease developed into a pandemic in March 2020 (WHO declared it a pandemic on March 11), the frequency value had exploded in the middle of January 2020. The increase in frequency occurred 11 days earlier than the increase in new cases. Figure 2b shows the Omicron VOC[5] (SARS-CoV-2 variant: B.1.1.529) pandemic, and Figure 2c depicts the spread of the Ebola virus (EBOV) around West Africa in 2014. Figure 2d shows when there were no severe epidemics. Table 3 summarizes the details of finding $k_{\max}$ and $\rho_{\max}$.

However, this detection via news frequency is limited by the lack of explainability. There is no further information other than the increase in pandemic-related news. For example, since the International Committee on Taxonomy of Viruses (ICTV) did not name it "Coronavirus 2" until February 11, 2020, we have no way to understand which disease raised the alarm shown in Figure 2a in January 2020. Our approach identifies its name in early January (Section 5.3). Moreover, we do not know why there was a modest surge in May 2014, as seen in Figure 2c, and in fact, this peak is not due to the EBOV outbreak (Section 5.3). We cannot persuade a decision-maker to implement preventative programs and the public to accept such policies unless we can explain the effect.

[5]Although Omicron variant of concern (VOC) is a mutant of COVID-19 virus, for convenience we distinguish the term COVID-19 as a virus that broke out in 2019–2020, and Omicron VOC as a virus that spread in the late 2021.

5. Theme Graph Clustering for Explainable Detection of Epidemics

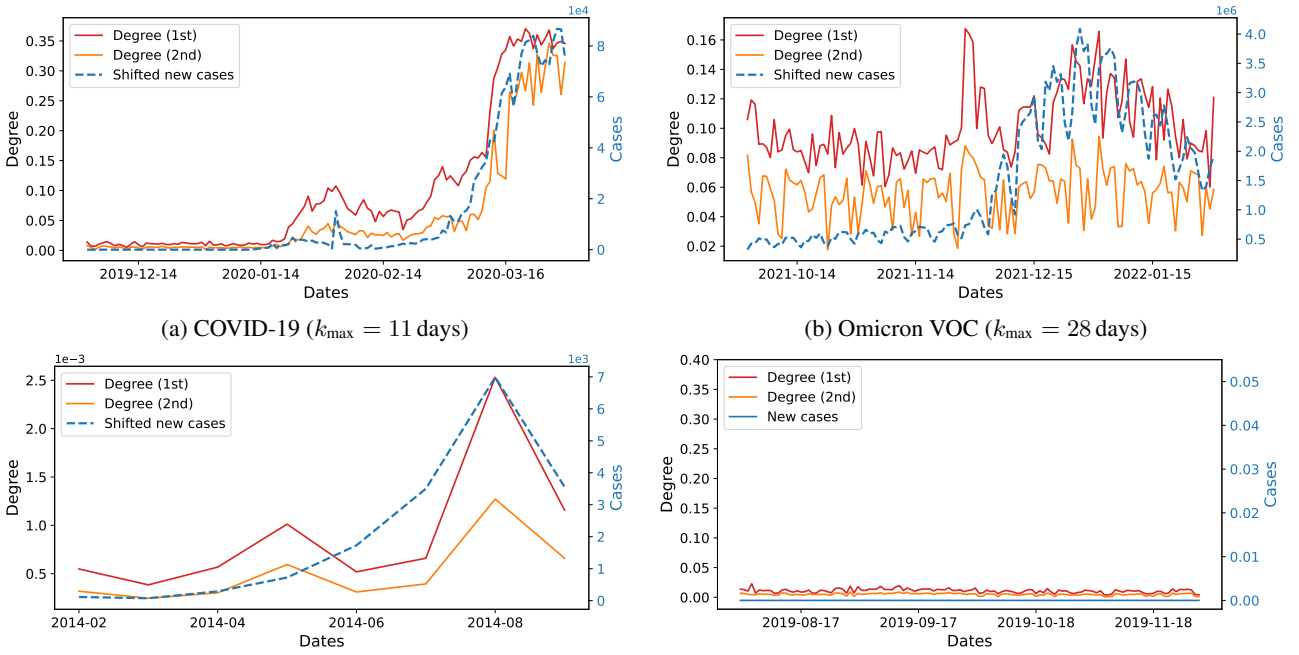
We cluster the theme nodes of a weighted graph using the Louvain algorithm presented in Section 3.3. Furthermore, we perform clustering twice in a bottom-up manner: (1) we perform clustering and find a cluster that contains the PANDEMICS node, then (2) repeat the clustering with a subgraph that contains only the intra-cluster nodes. This hierarchical clustering allows us to analyze two different levels of the clusters, which we call the 1st-level and 2nd-level clusters, respectively. We can detect the epidemics using these clusters, and more importantly, we can easily interpret the result.

5.1. Intra-Cluster Weighted Degree Centrality

Graph clustering enables us to find a few of the most relevant themes out of the 1,125 themes. Once we have found the cluster that contains the PANDEMICS node, we can quantify the associations between nodes within the cluster. We demonstrate that the pandemics degree as well as the pandemics frequency (in Section 4) effectively detects the epidemics. The pandemics degree is defined as the intra-cluster weighted degree centrality of the PANDEMICS node, *i.e.*, the sum of all the weights of the links between PANDEMICS and the nodes inside the cluster. The degree centrality informs us about the relative importance of a node; thus, a pandemics degree value implies the importance of the news items that are closely related to pandemics. We call it the 1st (or 1st-level) degree if it is measured within the 1st-level cluster and the 2nd (or 2nd-level) degree if measured within the 2nd-level cluster. For other centrality measures, such as closeness, betweenness, and PageRank (Page et al., 1999), refer to Appendix A.

Figure 3 shows the pandemics degrees in four different periods, and Table 4 summarizes their correlation evaluations by replacing the daily frequency in Eqs. (2) and (3) with the daily pandemics degree. When there is an ongoing epidemic, both degree measures detect the epidemics in the early phases of spread (Figures 3a–3c); when there is no existing epidemic, the pandemics degree does not raise a false alarm (Figure 3d). Specifically, the pandemics degree shows towering peaks before the epidemic spread. In particular, the pandemics degree for Omicron VOC (Figure 3b) shows a more significant peak than the pandemics frequency (Figure 2b) at the end of November 2021.

The 1st degree detects the COVID-19 epidemic 11 days earlier, the Omicron VOC epidemic 28 days earlier, and EBOV two months earlier. This result is better than or comparable to the frequency measure. Meanwhile, the $\rho_{\max}$ of the 2nd degree for Omicron VOC is very low, presumably due to the ongoing pandemic of Coronavirus and its variants. Although



(a) COVID-19 ($k_{\max} = 11$ days)

(b) Omicron VOC ($k_{\max} = 28$ days)

(c) Ebola virus ($k_{\max} = 2$ months)

(d) No epidemics

Figure 3. The pandemics degree measured over four different epidemiological periods. Degree (1st) and Degree (2nd) denote the inbound degrees of a PANDEMICS node within the 1st-level and the 2nd-level clusters, respectively. A dashed line indicates the graph of newly infected cases shifted by $k_{\max}$ days of the 1st degree (refer to Table 4).

Table 4. Correlation between the pandemics frequency/degree and the shifted new cases. The maximum correlation $\rho_{\max}$ occurs when the new cases graph is shifted by $k_{\max}$ days. Because of our theme clustering, we can gain explainability by using the pandemics degree.

Disease	Method	Explainable	$\rho_{\max}$	$k_{\max}$
COVID-19	Frequency	✗	97.14	**11**
	Degree (1st)	✓	96.37	**11**
	Degree (2nd)	✓	**97.79**	8
Omicron VOC	Frequency	✗	**56.66**	21
	Degree (1st)	✓	55.87	**28**
	Degree (2nd)	✓	26.08	23
Ebola virus	Frequency	✗	**86.73**	**2**
	Degree (1st)	✓	86.72	**2**
	Degree (2nd)	✓	**86.73**	**2**

Table 5. A list of themes with the highest degrees in the 2nd-level cluster. The themes that have not shown up in the top-10 list in the most recent seven days are highlighted. Top-10 themes from the 1st-level cluster are summarized in Appendix C.

Date (2020)	Top-10 Themes (ordered by intra-cluster degrees)
Jan 18	HEALTH·POINTSOFINTEREST·MEDICAL·EDUCATION·DISEASE DISEASES·NON·HEALTHCARE·SCIENCE·PHARMACEUTICALS
Jan 19	HEALTH·MEDICAL·DISEASES·DISEASE·NON HEALTHCARE·PHARMACEUTICALS·ORGANIZED·DRUGS·INJURY
Jan 20	HEALTH·MEDICAL·DISEASE·DISEASES·NON HEALTHCARE·PHARMACEUTICALS·ORGANIZED·DRUGS·INJURY
Jan 21	DISEASE·HEALTHCARE·**PANDEMICS**·**DEVELOPMENTORGS**·**AIRPORTS** **AIDGROUPS**·**HOLIDAY**·**PREVENTION**·**COMMUNICABLE**·**PANDEMIC**
Jan 22	HEALTH·MEDICAL·DISEASE·DISEASES·HEALTHCARE NON·PANDEMICS·DEVELOPMENTORGS·AIDGROUPS·PREVENTION
Jan 23	DISEASE·HEALTHCARE·PANDEMICS·DEVELOPMENTORGS·AIRPORTS AIDGROUPS·COMMUNICABLE·**DELAY**·HOLIDAY·PANDEMIC
Jan 24	HEALTHCARE·PANDEMICS·DEVELOPMENTORGS·AIRPORTS·AIDGROUPS HOLIDAY·COMMUNICABLE·**QUARANTINE**·DELAY·PREVENTION

the 1st degree might be preferred (based on the results in the COVID-19 and Omicron VOC cases), the 2nd degree is crucial for explainability. We can better figure out the most relevant themes through the 2nd-level clustering, as detailed in the following section.

5.2. Interpretation of Epidemic Detection

Theme Analysis. The theme graph clustering can explain the findings. We can determine which themes are pertinent to pandemics by examining the clusters obtained. Table 5 summarizes the 2nd-level cluster's themes with the highest intra-cluster degrees during the COVID-19 outbreak. To detect newly emerging themes in practice, specific rules must be established; we used an *unseen-for-one-week rule*. On January 21, 2020, when we can observe the rising degree value in Figure 3a, there were emerging themes including AIRPORTS, AIDGROUPS, COMMUNICABLE, DEVELOPMENTORGS, and PREVENTION. These extracted themes

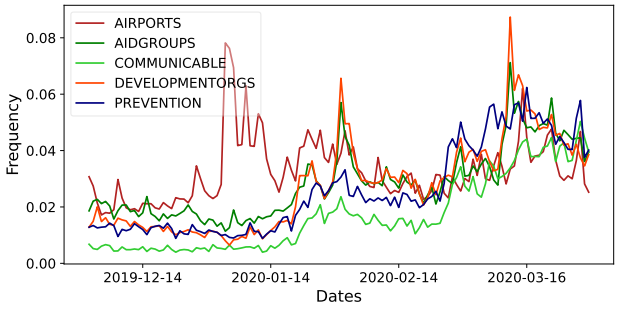


Figure 4. The news frequency based on the several themes that emerged in the 2nd-level cluster. Best seen in color.

(or keywords) can explain the epidemic alarm. While the 1st degree can also detect epidemics, it cannot provide such precise information because the 1st-level cluster contains several generic and irrelevant themes such as ETHNICITY or WORLDMAMMALS (refer to Appendix C).

News Retrieval. Moreover, we can measure the news frequency regarding the extracted themes. Figure 4 shows that the frequencies of corresponding themes also increased in January 2020. Among them, AIRPORTS shows a sharp increase beginning January 3. To retrieve the source of the information, we find the related news articles on this date through the following procedure:

1. Collect the news articles tagged with the subthemes of both PANDEMICS and AIRPORTS.
2. Sort the articles according to the *Average Tone* value that represents the overall tone of the article's contents.
3. Select the articles with the lowest *Average Tone* values.

From this, we list below the three lowest tone values and the corresponding articles (*Average Tone* value · News title · Source). Note that a lower tone value implies negative semantics in a news article (refer to Appendix E for details).

- ◇ -10.00 · PNEUMONIA OUTBREAK IN CHINA SPURS FEVER CHECKS FROM SINGAPORE TO TAIWAN · BLOOMBERGQUINT
- ◇ -9.64 · MOH TO SCREEN TRAVELLERS FROM WUHAN, CHINA FOLLOWING 'UNEXPLAINED' PNEUMONIA OUTBREAK · THE INDEPENDENT SINGAPORE
- ◇ -8.73 · MYSTERIOUS RESPIRATORY VIRUS STRIKES 44 PEOPLE IN CHINA · MSN NEWS

This information allows us to curate meaningful information about the disease and respond in advance.

Table 6. A list of Type D graph themes with the highest degrees in the 2nd-level cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted. In particular, the themes highlighted in red indicate subthemes of the DISEASE.

Date (2020)	Top-10 Themes (ordered by intra-cluster degrees)
Jan 07	DISEASE·HEALTHCARE·PANDEMICS·COMMUNICABLE·REPRODUCTIVE OUTBREAK·VACCINATION·CHILD·PREVENTION·INFECTION
Jan 08	HEALTHCARE·PANDEMICS·COMMUNICABLE·REPRODUCTIVE·VACCINATION CHILD·IMMUNIZATIONS·OUTBREAK·PREVENTION·INFECTION
Jan 09	HEALTHCARE·PANDEMICS·OUTBREAK·PREVENTION·COMMUNICABLE **PNEUMONIA**·INFECTION·REPRODUCTIVE·**CORONAVIRUS**·VACCINATION
Jan 10	HEALTHCARE·REPRODUCTIVE·VACCINATION·PANDEMICS·COMMUNICABLE CHILD·IMMUNIZATIONS·PREVENTION·INFECTION·**FEVER**
...	...
Jan 18	HEALTHCARE·PANDEMICS·OUTBREAK·**AIRPORTS**·CORONAVIRUS PNEUMONIA·**SARS**·COMMUNICABLE·**SYNDROME**·**SEVERE**
Jan 19	HEALTHCARE·PANDEMICS·OUTBREAK·CORONAVIRUS·PNEUMONIA AIRPORTS·SYNDROME·SARS·SEVERE·**FEVER**
Jan 20	HEALTHCARE·PANDEMICS·OUTBREAK·CORONAVIRUS·SARS PNEUMONIA·SYNDROME·SEVERE·PREVENTION·**HOLIDAY**

The 2nd-level clustering results during the Omicron VOC and EBOV periods are summarized in Appendix C. At the end of November 2021, when Omicron VOC began to spread, we observe newly emerging themes, including CHILD, DELAY, IMMUNIZATIONS, REPRODUCTIVE, and VACCINATION, which are highly related to a virus strike. Furthermore, in August 2014, when we detect that there will be an epidemiological spread, we observe themes that did not appear in July, including BAN, INEQUALITY, SLUMS, STRIKE, and UNREST. Themes like INEQUALITY and SLUMS may imply the relevance of poverty to the detected disease EBOV (Fallah et al., 2015).

5.3. DISEASE Subthemes Provide Better Explainability

As noted in Section 3.2, we can disassemble the DISEASE supertheme only. The **Type D** graph enhances the explainability of the clustering results with only a small increase in cost (Table 2). This provides a new graph containing specific information on the diseases, with 4,424 more nodes than the graph previously used in Section 5.1.

While the 1st degree approximately matches the shifted new cases ($\rho_{\max} = 95.88$, $k_{\max} = 11$), the 2nd-level clustering result can offer essential information about the pandemic details. Table 6 summarizes the top-10 themes in the 2nd-level cluster. On January 9, 2020, two new themes appeared in the cluster: PNEUMONIA and CORONAVIRUS. In the upcoming days, virus-related themes such as SARS and FEVER appeared. Given that the ICTV did not name this virus until February 11, this discovery could be used to inform people about the disease quickly.

The detailed theme lists for Omicron VOC and EBOV are

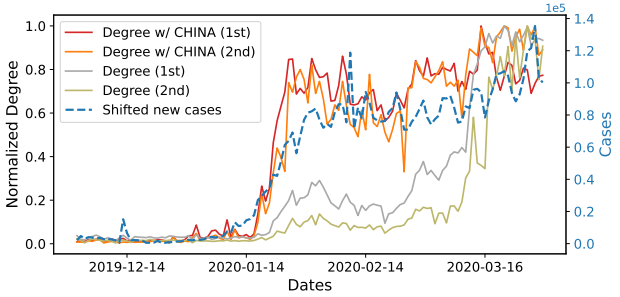
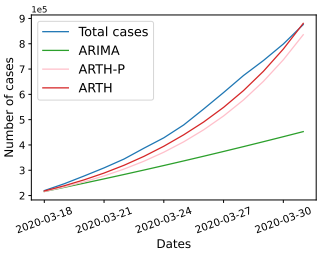


Figure 5. The pandemics degree measured when the regional information is given. For comparison, we also included the degree without the regional information, which is equivalent to the results in Figure 3a. A dashed line indicates the graph of newly infected COVID-19 cases shifted by 62 days.



Model	Metric	Error
ARIMA	MAPE	27.10
	WAPE	33.21
ARTH-P	MAPE	10.42
	WAPE	10.72
ARTH	MAPE	**6.02**
	WAPE	**5.94**

Figure 6. COVID-19 forecasting for total cases. Train: 02/05/2020–03/17/2020. Test: 03/18/2020–03/31/2020. MAPE stands for Mean Absolute Percentage Error, and WAPE stands for Weighted Average Percentage Error.

summarized in Appendix D. For Omicron VOC, we continue observing CORONAVIRUS with its related themes, such as AFFECT, PHARMACEUTICALS, PREVENTION, and OUTBREAK because there is a prevalent ongoing epidemic. Furthermore, our theme analysis reveals that the minor increase in May 2014 (Figure 3c) has little to do with the EBOV outbreak. During this period, the theme POLIO appears. In fact, a poliovirus had spread around East and Central Africa in 2013–2015, and WHO declared an epidemic of wild poliovirus on May 5, 2014 (ECDC, 2016). In August 2014, EBOLA emerges as a dominant theme in the 2nd-level cluster.

For a broader analysis of themes other than DISEASE, one can follow a similar procedure of generating new graphs and clustering. For example, we can generate a graph by disassembling only the FNCACT supertheme if we are interested in the occupational groups of people.

5.4. Epidemic Detection with Regional Information

In this section, we confirm that the earlier detection of epidemics is possible when regional information is available. The regional information about a specific disease can be readily obtained, and we can use the region to narrow down the relevant news. For instance, ProMED-mail (Yu & Madoff, 2004) sends daily reports of disease outbreaks around the world, with the specific region where the disease broke out.

One example of known origin is COVID-19, which was reported as an unknown respiratory disease that spread around Wuhan, China, in December 2019 (Bogoch et al., 2020; Chen et al., 2020; Wang et al., 2020; Wu et al., 2020). By using the regional information, we can consider the earlier outbreak in certain regions, for example, the two-month gap between the COVID-19 outbreak around China and its global surge. For COVID-19 epidemic detection, we select news tagged with CHINA in the *Locations* variable. Figure 5 shows the pandemics degree results. The 1st degree with the regional information achieved $\rho_{\max} = 95.00$ with $k_{\max} = 62$, significantly faster than the results without regional information.

6. Forecasting Cases with Theme Information

The extracted themes can be used to forecast the number of confirmed cases. We conducted a simple experiment to compare the forecasting performance with and without extracted themes. For the base model, we used the ARIMA (AutoRegressive Integrated Moving Average) model, which is often used as a baseline because of its simplicity and effectiveness. Inspired by ARGO (AutoRegression with GOogle search data) (Yang et al., 2015), which integrates search data to a regression model, we propose ARTH (AutoRegression with THeme). Rather than the Google search data, ARTH employs the frequency of themes extracted from the graph clustering.

We compare ARTH with ARIMA in Figure 6. Specifically, ARTH-P uses only the pandemics frequency as an input, whereas ARTH uses the frequency of (PANDEMICS + the five extracted themes in Figure 4). ARTH predicts the most accurately, while ARTH-P also outperforms ARIMA. This implies that including the information regarding the PANDEMICS theme helps forecast COVID-19 cases, with a considerably greater gain using additional extracted themes.

7. Conclusion

This study clarifies the feasibility of detecting epidemics before their spread without relying on data from confirmed cases. Our work is the first to use a news dataset for epidemic detection. We used the GKG news dataset, converted it into a graph, and for efficiency, integrated the subthemes. Furthermore, our approach detects epidemics and gives explainability through the use of 2nd-level cluster themes. In addition, dismantling a supertheme provides extra explainability regarding the supertheme. We believe that our work will inspire researchers to better detect epidemics without statistical data or domain knowledge and that our extracted themes can aid in forecasting patient numbers.

References

Abbasimehr, H. and Paki, R. Prediction of covid-19 confirmed cases combining deep learning methods and bayesian optimization. *Chaos, Solitons & Fractals*, 142: 110511, 2021.

Aiken, E. L., McGough, S. F., Majumder, M. S., Wachtel, G., Nguyen, A. T., Viboud, C., and Santillana, M. Real-time estimation of disease activity in emerging outbreaks using internet search information. *PLoS computational biology*, 16(8):e1008117, 2020.

Arik, S., Li, C.-L., Yoon, J., Sinha, R., Epshteyn, A., Le, L., Menon, V., Singh, S., Zhang, L., Nikoltchev, M., et al. Interpretable sequence learning for covid-19 forecasting. *Advances in Neural Information Processing Systems*, 33: 18807–18818, 2020.

Benson, A. R., Gleich, D. F., and Leskovec, J. Higher-order organization of complex networks. *Science*, 353(6295): 163–166, 2016.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Bogoch, I. I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M. U., and Khan, K. Pneumonia of unknown aetiology in wuhan, china: potential for international spread via commercial air travel. *Journal of travel medicine*, 27 (2):taaa008, 2020.

Bosa, I., Castelli, A., Castelli, M., Ciani, O., Compagni, A., Galizzi, M. M., Garofano, M., Ghislandi, S., Giannoni, M., Marini, G., et al. Response to covid-19: was italy (un) prepared? *Health Economics, Policy and Law*, 17(1): 1–13, 2022.

Chakraborty, A. and Bose, S. Around the world in 60 days: an exploratory study of impact of covid-19 on online global news sentiment. *Journal of computational social science*, 3(2):367–400, 2020.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395(10223):507–513, 2020.

Chimmula, V. K. R. and Zhang, L. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, 135:109864, 2020.

Chodrow, P. S., Veldt, N., and Benson, A. R. Generative hypergraph clustering: From blockmodels to modularity. *Science Advances*, 7(28):eabh1303, 2021.

Deng, Q. et al. Dynamics and development of the covid-19 epidemic in the united states: a compartmental model enhanced with deep learning techniques. *Journal of Medical Internet Research*, 22(8):e21173, 2020.

Du, B., Zhao, Z., Zhao, J., Yu, L., Sun, L., and Lv, W. Modelling the epidemic dynamics of covid-19 with consideration of human mobility. *International Journal of Data Science and Analytics*, 12(4):369–382, 2021.

ECDC. Annual epidemiological report 2016 – poliomyelitis. Technical report, European Centre for Disease Prevention and Control, 2016. URL https://www.ecdc.europa.eu/en/publications-data/polio-annual-epidemiological-report-2016-2014-data.

Fallah, M. P., Skrip, L. A., Gertler, S., Yamin, D., and Galvani, A. P. Quantifying poverty as a driver of ebola transmission. *PLoS neglected tropical diseases*, 9(12): e0004260, 2015.

Feng, R., Hu, Q., and Jiang, Y. Unknown disease outbreaks detection: A pilot study on feature-based knowledge representation and reasoning model. *Frontiers in Public Health*, 9:535, 2021.

Fu, K.-w. and Zhu, Y. Did the world overlook the media's early warning of covid-19? *Journal of Risk Research*, 23 (7-8):1047–1051, 2020.

Galla, D. and Burke, J. Predicting social unrest using gdelt. In *International conference on machine learning and data mining in pattern recognition*, pp. 103–116. Springer, 2018.

Hao, Q., Chen, L., Xu, F., and Li, Y. Understanding the urban pandemic spreading of covid-19 with real world mobility data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3485–3492, 2020.

Hashimoto, S., Murakami, Y., Taniguchi, K., and Nagai, M. Detection of epidemics in their early stage through infectious disease surveillance. *International journal of epidemiology*, 29(5):905–910, 2000.

Jacobs, G., Lefever, E., and Hoste, V. Economic event detection in company-specific news text. In *1st Workshop on Economics and Natural Language Processing (ECONLP) at Meeting of the Association-for-Computational-Linguistics (ACL)*, pp. 1–10. Association for Computational Linguistics (ACL), 2018.

Jakel, T. Using sentiment data from the global database for events, language and tone (gdelt) to predict short-term stock price developments. B.S. thesis, University of Twente, 2019.

Jin, X., Wang, Y.-X., and Yan, X. Inter-series attention model for covid-19 forecasting. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 495–503. SIAM, 2021.

Kim, M., Kang, J., Kim, D., Song, H., Min, H., Nam, Y., Park, D., and Lee, J.-G. Hi-covidnet: deep learning approach to predict inbound covid-19 patients and case study in south korea. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3466–3473, 2020.

Korolev, I. Identification and estimation of the SEIRD epidemic model for COVID-19. *Journal of econometrics*, 220(1):63–85, 2021.

Krawczyk, K., Chelkowski, T., Laydon, D. J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., et al. Quantifying online news media coverage of the covid-19 pandemic: Text mining study and resource. *Journal of Medical Internet Research*, 23 (6):e28253, 2021.

Lucas, B., Elliot, B., and Landman, T. Online information search during COVID-19. *arXiv preprint arXiv:2004.07183*, 2020.

Menda, K., Laird, L., Kochenderfer, M. J., and Caceres, R. S. Explaining covid-19 outbreaks with reactive seird models. *Scientific Reports*, 11(1):1–12, 2021.

Ozili, P. K. and Arun, T. Spillover of covid-19: impact on the global economy. *Available at SSRN 3562570*, 2020.

Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Qian, Z., Alaa, A. M., and van der Schaar, M. When and how to lift the lockdown? global covid-19 scenario analysis and policy assessment using compartmental gaussian processes. *Advances in Neural Information Processing Systems*, 33:10729–10740, 2020.

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., and Roser, M. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

Saz-Carranza, A., Maturana, P., and Quer, X. The empirical use of gedlt big data in academic research. Technical report, Global Governance and the European Union, 2020. URL https://www.globe-project.eu/the-empirical-use-of-gdelt-big-data-in-academic-research_13809.pdf.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., and Roychowdhury, V. Conspiracy in the time of corona: Automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317, 2020.

Vega, R., Flores, L., and Greiner, R. Simlr: Machine learning inside the sir model for covid-19 forecasting. *Forecasting*, 4(1):72–94, 2022.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama*, 323(11):1061–1069, 2020.

Wu, J. T., Leung, K., and Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395(10225):689–697, 2020.

Yang, S., Santillana, M., and Kou, S. C. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112 (47):14473–14478, 2015.

Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of thoracic disease*, 12(3):165, 2020.

Yu, V. L. and Madoff, L. C. Promed-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2):227–232, 2004.

Zhou, D., Huang, J., and Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.

Appendix

A. Epidemic Detection with Different Centrality Measures

In our experiments in Section 5, we used the degree centrality for detecting the epidemic. Specifically, we measured the intra-cluster degree centrality of the PANDEMICS node with the clusters obtained. In addition to the degree centrality, there are several other centrality measures that evaluate the node characteristics or properties. In this section, we consider three measures that are most commonly used: closeness centrality, betweenness centrality, and PageRank centrality (Page et al., 1999). We briefly explain each centrality below.

- **Closeness centrality** assigns a score to each node based on how much it is close to all other nodes in a network. It is calculated by the inverse of the average shortest path to all other nodes.
- **Betweenness centrality** counts the number of shortest paths between any pair of nodes in a network that travel through the node.
- **PageRank centrality** is a variant of eigenvector centrality and assigns a score to each node that represents the node importance, based on their connections and weights. It is originally designed for directed graphs.

Each centrality measure within the cluster is computed in the same way the pandemics degree is computed. Figure 7 shows the results of three distinct centrality measures. These measures exhibit significant noise and cannot effectively detect epidemics. The closeness and betweenness are concerned with the node's connectivity, whereas the degree and PageRank are concerned with the node's prominence (or influence) in the network. We discover that prominence is an excellent indicator of epidemics. While the PageRank is as excellent as the degree centrality in detecting the epidemic (see Figure 7 right), it exhibits some noise even before the outbreak. In contrast, the degree does not (refer to Figures 3a and 3d). We suppose that PageRank does not perform optimally in our densely connected undirected graphs. A . We leave it as future work to study deeper with other centrality measures.

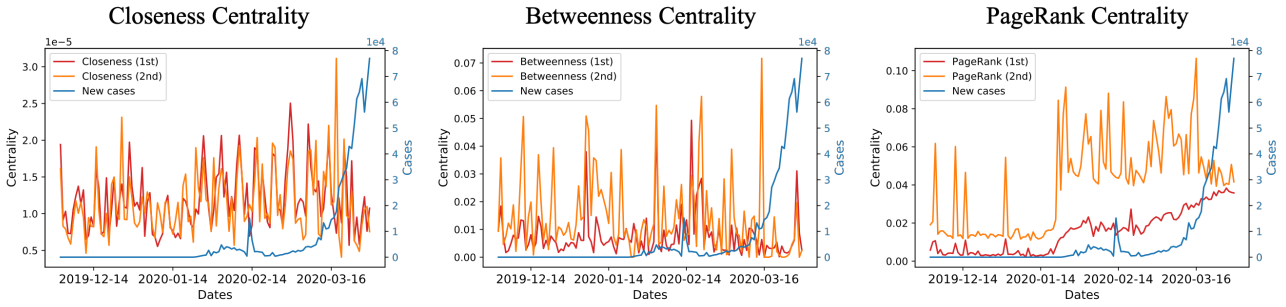


Figure 7. The three different centrality measures of the PANDEMICS node for COVID-19 detection. We use closeness centrality, betweenness centrality, and PageRank centrality. The evaluation period is the same as that in Figure 3a.

B. Pandemics Degree Plot with Type D Graph (Section 5.3)

The **Type D** graph effectively detects disease-related themes, as seen in the theme analysis in Section 5.3. For a thorough analysis, we provide the pandemics degree plot with the **Type D** graph in Figure 8. The 1st degree in the **Type D** graph also exhibits a high correlation to the newly infected cases of COVID-19, while the 2nd degree exhibits a large variance in March 2020. For the 1st degree, $\rho_{\max} = 95.88$, $k_{\max} = 11$, and for the 2nd degree, $\rho_{\max} = 87.31$, $k_{\max} = 7$.

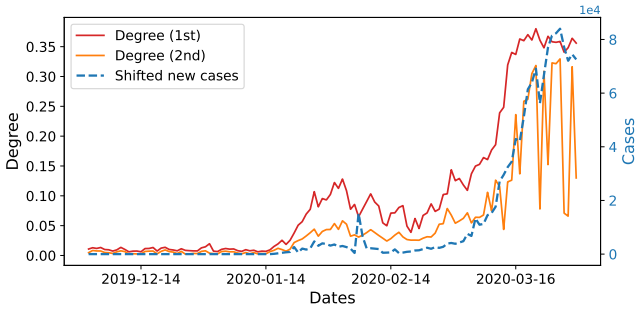


Figure 8. The pandemics degree measured with the Type D graph. Degree (1st) and Degree (2nd) denote the inbound degrees of a PANDEMICS node within the 1st-level and the 2nd-level clusters, respectively. A dashed line indicates the graph of newly infected COVID-19 cases shifted by 11 days.

C. Additional Theme Analyses of Type R Graph

In this section, we describe additional theme analyses during a specific period for each epidemic using the **Type R** graph.

C.1. COVID-19 (1st-Level Cluster)

Table 7 summarizes the top-10 themes in the 1st-level cluster during 01/18/2020–01/24/2020, the same period as Table 5. Newly emerging themes, *i.e.*, DISEASE, TRANSPORT, WORLDMAMMALS, and DISEASES, only appear on January 18, 2020. In comparison to Table 5, we can find no significant change in the 1st-level cluster's themes. In addition, the 1st-level cluster contains generic themes such as CRISISLEX, FNCACT, and KILL and irrelevant themes such as ETHNICITY, FORESTS, and WORLDMAMMALS.

Table 7. A list of themes with the highest degrees in the **1st-level** cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted.

Date (2020)	Top-10 Themes (ordered by intra-cluster degrees)
Jan 18	HEALTH·POINTSOFINTEREST·DISASTER·FORESTS·EDUCATION·MEDICAL·**DISEASE·TRANSPORT·WORLDMAMMALS·DISEASES**
Jan 19	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·DISASTER·FORESTS·POINTSOFINTEREST·HEALTH·WORLDLANGUAGES·EDUCATION
Jan 20	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·DISASTER·POINTSOFINTEREST·WORLDLANGUAGES·MEDICAL·KILL
Jan 21	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·POINTSOFINTEREST·WORLDLANGUAGES·DISASTER·MEDICAL·EDUCATION
Jan 22	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·POINTSOFINTEREST·WORLDLANGUAGES·DISASTER·MEDICAL·EDUCATION
Jan 23	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·POINTSOFINTEREST·DISASTER·WORLDLANGUAGES·MEDICAL·KILL
Jan 24	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·POINTSOFINTEREST·WORLDLANGUAGES·DISASTER·MEDICAL·KILL

C.2. Omicron VOC

Table 8 summarizes the top-10 themes in the 2nd-level cluster during 11/17/2021–11/30/2021. This is the period when we detect that there will be an epidemiological spread, as shown in Figure 3b. New themes appeared beginning November 20, 2021, such as CHILD, DELAY, IMMUNIZATION, REPRODUCTIVE, and VACCINATION. Since COVID-19 was widely spreading at the time, themes like DELAY, REPRODUCTIVE, and VACCINATION may reflect Omicron VOC's relevance to COVID-19.

Table 8. A list of themes with the highest degrees in the 2nd-level cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted. The dates with less than ten themes indicate that the cluster size is smaller than ten.

Date (2021)	Top-10 Themes (ordered by intra-cluster degrees)
Nov 17	FNCACT·HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC
Nov 18	FNCACT·HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC
Nov 19	FNCACT·HEALTH·PUBLIC·DISEASE·POINTOFINTEREST·HEALTHCARE·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC
Nov 20	HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·**EDUCATIONAL**
Nov 21	HEALTH·PUBLIC·DISEASE·POINTSOFINTEREST·HEALTHCARE·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·**DELAY**
Nov 22	HEALTH·HEALTHCARE·DISEASE·MEDICAL·PANDEMICS·PANDEMIC·**VACCINATION**·**REPRODUCTIVE**·**CHILD**·**IMMUNIZATIONS**
Nov 23	FNCACT·HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC
Nov 24	HEALTHCARE·PANDEMICS·PANDEMIC·VACCINATION·REPRODUCTIVE·CHILD·IMMUNIZATIONS·**PREVENTION**·**COMMUNICABLE**·**QUARANTINE**
Nov 25	HEALTHCARE·PANDEMICS·PANDEMIC·**HAZMAT**·**UNGOVERNED**·**MUNICIPAL**
Nov 26	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·POINTSOFINTEREST·PANDEMIC·EDUCATION·**SCIENCE**
Nov 27	HEALTH·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·VACCINATION·PANDEMIC·REPRODUCTIVE·CHILD·IMMUNIZATION
Nov 28	HEALTH·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·VACCINATION·PANDEMIC·REPRODUCTIVE·CHILD·IMMUNIZATION
Nov 29	FNCACT·HEALTH·**CRISISLEX**·PUBLIC·**POLICY**·DISEASE·ETHNICITY·MEDICAL·HEALTHCARE·**CAT**
Nov 30	FNCACT·HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·POINTSOFINTEREST·EDUCATION·PANDEMICS·PANDEMIC

C.3. Ebola Virus

Table 9 summarizes the top-10 themes in the 2nd-level cluster during 08/01/2014–08/07/2014 and 08/16/2014–08/22/2014. We examine these periods since we identify a large peak in August 2014, as shown in Figure 3c. The themes, which rarely appeared in July 2014, continually appear in August 2014, *e.g.*, POVERTY, SANITATION, and SLUMS. Note that in 2014, the subtheme WB-2167-PANDEMICS did not exist, so we used PANDEMIC as a supertheme (the corresponding subtheme is HEALTH-PANDEMIC).

Table 9. A list of themes with the highest degrees in the 2nd-level cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted. The dates with less than ten themes indicate that the cluster size is smaller than ten.

Date (2014)	Top-10 Themes (ordered by intra-cluster degrees)
Aug 01	TURMOIL·OF·VACCINATION·PANDEMIC·SICKENED
Aug 02	HEALTH·MEDICAL·**EDUCATION**·DISEASE·SICKENED·VACCINATION·**TRAFFIC**·DISABILITY·**UNREST**·PANDEMIC
Aug 03	HEALTH·MEDICAL·DISEASE·**BAN**·VACCINATION·DISABILITY·SANITATION·SICKENED·**STRIKE**·SEXTRANSDISEASE
Aug 04	HEALTH·MEDICAL·DISEASE·SECURITY·TURMOIL·VACCINATION·OF·DISABILITY·SICKENED·SANITATION
Aug 05	HEALTH·MEDICAL·DISEASE·VACCINATION·SICKENED·SEXTRANSDISEASE·**OWNERSHIP**·PANDEMIC·**INEQUALITY**
Aug 06	TURMOIL·OF·PANDEMIC
Aug 07	HEALTH·MEDICAL·DISEASE·TURMOIL·OF·VACCINATION·PANDEMIC·SICKENED·SEXTRANSDISEASE·**PEACE**
...	...
Aug 16	HEALTH·MEDICAL·DISEASE·VACCINATION·PANDEMIC·SEXTRANSIDEASE·SICKENED·WORK·**MARKET**·**UNGOVERNED**
Aug 17	HEALTH·MEDICAL·DISEASE·**BAN**·VACCINATION·**SLUMS**·DISABILITY·SANITATION·PANDEMIC·SEXTRANSDISEASE
Aug 18	HEALTH·MEDICAL·DISEASE·VACCINATION·SLUMS·**ATTACK**·PANDEMIC·SICKENED·SEXTRANSDISEASE
Aug 19	HEALTH·MEDICAL·DISEASE·SECURITY·SLUMS·SANITATION·SICKENED·VACCINATION·SHORTAGE·PANDEMIC
Aug 20	HEALTH·MEDICAL·DISEASE·SECURITY·**CURFEW**·SLUMS·SHORTAGE·POVERTY·SICKENED·VACCINATION
Aug 21	HEALTH·MEDICAL·DISEASE·VACCINATION·PANDEMIC·SEXTRANSDISEASE
Aug 22	HEALTH·MEDICAL·DISEASE·VACCINATION·SANITATION·SEXTRANSDISEASE·PANDEMIC

C.4. Complete Themes List in the 2nd-Level Cluster

Table 10 summarizes the entire 2nd-level cluster's themes. The cluster size varies by date because the Louvain algorithm does not fix the community size. In Table 5, we observed new themes on January 21, 2020, hence in Table 10, we list the themes before and after a month of that date. Note that the cluster size on January 21 diminishes because the edge weights between PANDEMICS and highly relevant themes are rather large compared to the edge weights with other irrelevant themes.

Table 10. A list of the entire themes in the 2nd-level cluster. **Size** indicates the number of nodes in a cluster.

Date	Themes (ordered by intra-cluster degrees)	Size
Dec 21, 2019	HEALTH·MEDICAL·DISEASES·DISEASE·NON·PHARMACEUTICALS·ORGANIZED·DRUGS·HEALTHCARE·INJURY·CANCER·ILLEGAL·FOOD·PREVENTION·PANDEMICS·CHRONICDISEASE·MENTAL·NUTRITION·REPRODUCTIVE·ALCOHOL·COMMUNICABLE·ELDERLY·VACCINATION·DEMOGRAPHIC·CHILD·DISABILITY·IMMUNIZATIONS·EMERGENCY·SUPPLEMENTS·HEART·DIABETES·NUTRITIONAL·THERAPEUTIC·OBESITY·EMERGENCYROOM·INFLUENZA·AGING·CONTINUUM·NURSING·HYPERTENSION·LIFE·PRIMARY·PANDEMIC·EBOLA·SICKENED·FAMILY·SEXTRANSDISEASE·VULNERABLE·GENERIC·BREASTFEEDING·TOBACCO·CONTRACEPTIVES·DENSITY·TUBERCULOSIS·MALARIA·GERIATRICS·ORPHANS·HEALTHY·STUNTING·MIDWIVES·PREMATURE·SECONDARY·VULNERABILITY·AGINGPOPULATION·COMMUNITY·PARENT·ANTENATAL·PREVENTIVE·PALLIATIVE·STI·MATERIAL·ZINC·UNIVERSAL·COMBATANTS·COMBATANT·HIGH·ZOONOTIC·STREET·HIV·NEGLECTED·MICRONUTRIENTS·HOSPITAL·EPIDEMIOLOGY·INDUSTRIALACCIDENT·FUNGUS·COUNTERFEITFOOD	86
Jan 21, 2020	DISEASE·HEALTHCARE·PANDEMICS·DEVELOPMENTORGS·AIRPORTS·AIDGROUPS·HOLIDAY·PREVENTION·COMMUNICABLE·PANDEMIC·REPRODUCTIVE·TOURISM·VACCINATION·QUARANTINE·CHILD·IMMUNIZATIONS·SICKENED·SURVEILLANCE·RAILWAYS·EBOLA·INFLUENZA·SEXTRANSDISEASE·FAMILY·VULNERABILITY·TUBERCULOSIS·CONTRACEPTIVES·STI·HAZMAT·MALARIA·MIDWIVES·ANTENATAL·ZOONOTIC·HIV·PLANTDISEASE·ADOLESCENT·PARENTS·WASTE	37
Feb 21, 2020	HEALTH·DISEASE·MEDICAL·DISEASES·HEALTHCARE·PANDEMICS·NON·QUARANTINE·DEVELOPMENTORGS·PREVENTION·AIDGROUPS·FOOD·COMMUNICABLE·INJURY·REPRODUCTIVE·CHILD·VACCINATION·EVACUATION·IMMUNIZATIONS·CANCER·ELDERLY·CHRONICDISEASE·PANDEMIC·MENTAL·INFLUENZA·NUTRITIONAL·SICKENED·THERAPEUTIC·HEART·DIABETES·REPATRIATION·UNREST·EBOLA·HYPERTENSION·OBESITY·SEXTRANSDISEASE·FAMILY·CONTINUUM·EMERGENCYROOM·STONETHROWING·GENERIC·TUBERCULOSIS·PRIMARY·BREASTFEEDING·NURSING·CONTRACEPTIVES·MALARIA·CLOSINGBORDER·STI·HIGH·TOBACCO·HAZMAT·MIDWIVES·GERIATRICS·FIELDHOSPITAL·UNIVERSAL·PREMATURE·DRINKING·PALLIATIVE·NEGLECTED·ESSENTIAL·ADOLESCENT·MATERNAL·PLANTDISEASE·PREVENTIVE·MALE·EPIDEMIOLOGY·BURDEN·PERSISTENT·SEDENTARY·ZOONOTIC·COMBATANTS·COMBATANT·HOSPITAL·HIV·FUNGUS·HELMINTH·COCREATION·EFFECTIVE·EXPERT	80

D. Additional Theme Analyses of Type D Graph

In this section, we describe additional theme analyses during a specific period for each epidemic using the **Type D** graph.

D.1. COVID-19 (1st-Level Cluster)

Table 11 summarizes the top-10 themes in the 1st-level cluster during 01/07/2020–01/10/2020 and 01/18/2020–01/20/2020, the same period as Table 6. However, in comparison to Table 6, there is no newly emerging theme in this period. In addition, the 1st-level cluster contains generic themes such as HEALTH and POINTSOFINTEREST and irrelevant themes such as DRUGS, WATER, and WORLDMAMMALS. Moreover, we cannot observe subthemes of the DISEASE supertheme, although we used the **Type D** graph.

Table 11. A list of Type D graph themes with the highest degrees in the **1st-level** cluster.

Date (2020)	Top-10 Themes (ordered by intra-cluster degrees)
Jan 07	HEALTH·POINTSOFINTEREST·MEDICAL·DISEASE·EDUCATION·DISEASES·NON·PHARMACEUTICALS·HEALTHCARE·SCIENCE
Jan 08	HEALTH·MEDICAL·DISEASE·DISEASES·NON·WORLDMAMMALS·AFFECT·PHARMACEUTICALS·HEALTHCARE·ORGANIZED
Jan 09	HEALTH·MEDICAL·DISEASE·DISEASES·NON·HEALTHCARE·PHARMACEUTICALS·SCIENCE·ORGANIZED·DRUGS
Jan 10	HEALTH·FORESTS·MEDICAL·DISEASE·DISEASES·NON·WORLDMAMMALS·AFFECT·AGRICULTURE·WATER
...	...
Jan 18	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·POINTSOFINTEREST·HEALTH·DISASTER·FORESTS·WORLDLANGUAGES·EDUCATION
Jan 19	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·DISASTER·POINTSOFINTEREST·WORLDLANGUAGES·MEDICAL·KILL
Jan 20	FNCACT·CRISISLEX·CRISISLEXREC·ETHNICITY·HEALTH·DISASTER·POINTSOFINTEREST·WORLDLANGUAGES·MEDICAL·KILL

D.2. Omicron VOC

Table 12 summarizes the top-10 themes in the 2nd-level cluster during 11/17/2021–11/30/2021. This is the period when we detect that there will be an epidemiological spread, as shown in Figure 3b. New themes appeared, such as AFFECT, DELAY, PHARMACEUTICALS, and PREVENTION. In particular, the themes related to COVID-19, such as CORONAVIRUS, continually appear in this period. This emphasizes the relevance between COVID-19 and Omicron VOC, and better explains the effect than the **Type R** graph does (refer to Table 8).

Table 12. A list of Type D graph themes with the highest degrees in the 2nd-level cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted. In particular, the themes highlighted in red indicate subthemes of the DISEASE supertheme.

Date (2021)	Top-10 Themes (ordered by intra-cluster degrees)
Nov 17	HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·CORONAVIRUS
Nov 18	HEALTH·HEALTHCARE·DISEASE·MEDICAL·PANDEMICS·PANDEMIC·VACCINATION·REPRODUCTIVE·CHILD·IMMUNIZATIONS
Nov 19	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·POINTSOFINTEREST·EDUCATION·PANDEMICS·PANDEMIC·DISEASES
Nov 20	HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·CORONAVIRUS
Nov 21	HEALTH·PUBLIC·DISEASE·POINTSOFINTEREST·HEALTHCARE·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·**AFFECT**
Nov 22	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·PANDEMIC·DISEASES·CORONAVIRUS·**HOLIDAY**
Nov 23	HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC·CORONAVIRUS
Nov 24	HEALTH·PUBLIC·DISEASE·HEALTHCARE·PANDEMICS·PANDEMIC·MEDICAL·CORONAVIRUS·HOLIDAY·**DELAY**
Nov 25	FNCACT·HEALTH·PUBLIC·DISEASE·HEALTHCARE·POINTSOFINTEREST·MEDICAL·EDUCATION·PANDEMICS·PANDEMIC
Nov 26	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·POINTSOFINTEREST·PANDEMIC·EDUCATION·**SCIENCE**
Nov 27	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·POINTSOFINTEREST·**CATS**·PANDEMICS·EDUCATION·PANDEMIC
Nov 28	HEALTH·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·PANDEMIC·DISEASES·SCIENCE·**PHARMACEUTICALS**·**PREVENTION**
Nov 29	HEALTH·DISEASE·PUBLIC·HEALTHCARE·MEDICAL·CATS·PANDEMICS·PANDEMIC·DISEASES·**AFFECT**
Nov 30	HEALTH·PUBLIC·DISEASE·HEALTHCARE·MEDICAL·PANDEMICS·PANDEMIC·DELAY·HOLIDAY·**OUTBREAK**

D.3. Ebola Virus

Table 13 summarizes the top-10 themes in the 2nd-level cluster during 05/01/2014–05/07/2014 and 08/01/2014–08/07/2014. We examine these periods since we identify two peaks in May 2014 and August 2014, as shown in Figure 3c. In comparison to Table 9, we can observe EBOLA in August, implying that the peak in August 2014 is attributed to the EBOV spread. In addition, during 05/02/2014–05/07/2014, we can detect themes relevant to a poliovirus, such as BACTERIA, MALARIA, and POLIO. This indicates that the peak in May 2014 is due to the poliovirus. When using the **Type R** graph, these themes are not observable and we are unable to estimate which disease will spread.

Table 13. A list of Type D graph themes with the highest degrees in the 2nd-level cluster. The themes that did not appear in the top-10 list in the most recent seven days are highlighted. In particular, the themes highlighted in red indicate subthemes of the DISEASE supertheme.

Date (2014)	Top-10 Themes (ordered by intra-cluster degrees)
May 01	HEALTH·MEDICAL·DISEASE·SCIENCE·CANCER·VACCINATION·INFECTION·OUTBREAK·DISABILITY·FEVER
May 02	SEXTRANSDISEASE·BACTERIA·MALARIA·PATHOGENS·SYPHILIS·SMALLPOX·ANTHRAX·PANDEMIC·ANTIBIOTIC·BACTERIAL
May 03	HEALTH·MEDICAL·DISEASE·MERS·TRANSPORT·VACCINATION·INFECTION·OUTBREAK·DISABILITY·FEVER
May 04	HEALTH·MEDICAL·DISEASE·VACCINATION·INFECTION·FEVER·MERS·CANCER·OUTBREAK·CORONAVIRUS
May 05	VACCINATION·POLIO·PANDEMIC·OUTBREAK·SMALLPOX·REFUGEES·SEXTRANSDISEASE·SURVEILLANCE·HEPATITIS·TREASON
May 06	DISEASE·VACCINATION·CANCER·OUTBREAK·POLIO·INFECTION·PANDEMIC·FEVER·SURVEILLANCE·SEXTRANSDISEASE
May 07	VACCINATION·INFECTION·OUTBREAK·FLU·INFECTIOUS·MERS·PNEUMONIA·CORONAVIRUS·PANDEMIC·POLIO
...	...
Aug 01	DISEASE·EBOLA·OUTBREAK·INFECTIOUS·FEVER·INFECTION·VACCINATION·CONTAGIOUS·HEMORRHAGIC·FLU
Aug 02	DISEASE·EBOLA·OUTBREAK·INFECTION·INFECTIOUS·FEVER·SICKENED·CONTAGIOUS·FLU·HEMORRHAGIC
Aug 03	DISEASE·EBOLA·OUTBREAK·INFECTION·INFECTIOUS·FEVER·HEMORRHAGIC·SARS·SICKENED·CONTAGIOUS
Aug 04	DISEASE·EBOLA·OUTBREAK·FEVER·INFECTION·INFECTIOUS·HEMORRHAGIC·VACCINATION·SICKENED·INFLUENZA
Aug 05	EBOLA·OUTBREAK·FEVER·INFECTIOUS·VACCINATION·INFECTION·HEMORRHAGIC·SICKENED·FLU·SARS
Aug 06	DISEASE·EBOLA·OUTBREAK·FEVER·INFECTION·VACCINATION·HEMORRHAGIC·CONTAGIOUS·SICKENED·INFECTIOUS
Aug 07	DISEASE·EBOLA·OUTBREAK·FEVER·INFECTION·INFECTIOUS·TURMOIL·OF·HEMORRHAGIC·VACCINATION

E. Using Tone Variable in GKG Dataset

The *Tone* variable in GKG dataset contains six emotional dimensions: *Average Tone*, *Positive Score*, *Negative Score*, *Polarity*, *Activity Reference Density*, and *Self/Group Reference Density*. The dimensions are explained in detail below.

- ***Average Tone***. The average of emotional connotations of the article's contents; (*Positive Score*) − (*Negative Score*). Its range is -100 to +100.
- ***Positive Score.*** The percentage of words with positive emotional connotations, in a range 0 to +100.
- ***Negative Score***. The percentage of words with negative emotional connotations, in a range 0 to +100.
- ***Polarity***. The percentage of words with emotional connotations. A low *Average Tone* value with a high *Polarity* value indicates that there are a lot of negative and positive emotional words but with similar amounts.
- ***Activity Reference Density***. The percentage of words that provide basic proxy of the overall activeness.
- ***Self/Group Reference Density***. The percentage of pronouns, capturing self-references and group-based discourse.

These *Tone* values allow us to search negative (or positive) and semantically particular news easily. In Section 5.2, we used the *Average Tone* value for **News Retrieval**. Besides, we are interested in using the tone values to detect the epidemics. To this end, after the theme graph clustering, we capture the 2nd-level cluster's tone values and the PANDEMICS theme's tone values.

For each date, we select the news that contains the 2nd-level cluster's themes, and then compute two normalized metrics: *Normalized Tone* and *Normalized Tone via Entire News*. Also, we select the news that contains the PANDEMICS theme and compute two normalized metrics: *Normalized Pandemics Tone* and *Normalized Pandemics Tone via Entire News*. These metrics are explained in detail below. Note that the tone value can be any of the six emotional dimensions.

- ***Normalized Tone (NT)*** is calculated by the sum of tone values of every news that contains the 2nd-level cluster's themes. We normalize this with maximum (or minimum if the sum is negative) tone value among the news.

$$NT = \frac{ST}{MT} \tag{4}$$

  where ST denotes the sum of tone values, and MT denotes the maximum (minimum) tone value. Since *Average Tone* can have negative values, we use the minimum value as MT only for *Average Tone*, and maximum value for other tone values.

- ***Normalized Tone via Entire News (NTE)*** sums the tone values in the same way as NT. However, we normalize the sum with average tone value of the corresponding news.

$$NTE = \frac{ST}{AT} \tag{5}$$

  where AT denotes the average tone value of the entire news.

- ***Normalized Pandemics Tone (NPT)*** is calculated by the sum of tone values of every news that contains the PANDEMICS theme. We normalize this with maximum (or minimum if the sum is negative) tone value among the news.

$$NPT = \frac{PT}{MT} \tag{6}$$

  where PT denotes the sum of tone values of the news containing PANDEMICS.

- ***Normalized Pandemics Tone via Entire News (NPTE)*** sums the tone values in the same way as NPT. However, we normalize the sum with average tone value of the corresponding news.

$$NPTE = \frac{PT}{AT} \tag{7}$$

Figure 9 shows four normalized metrics plotted against each emotional dimension in the *Tone* variable (*i.e.*, *Average Tone*, *Positive Score*, *Negative Score*, *Polarity*, *Activity Reference Density*, and *Self/Group Reference Density*). However, when compared to using the pandemics degree (refer to Figure 3), using the tone information for detecting the epidemic does not produce significantly superior results.

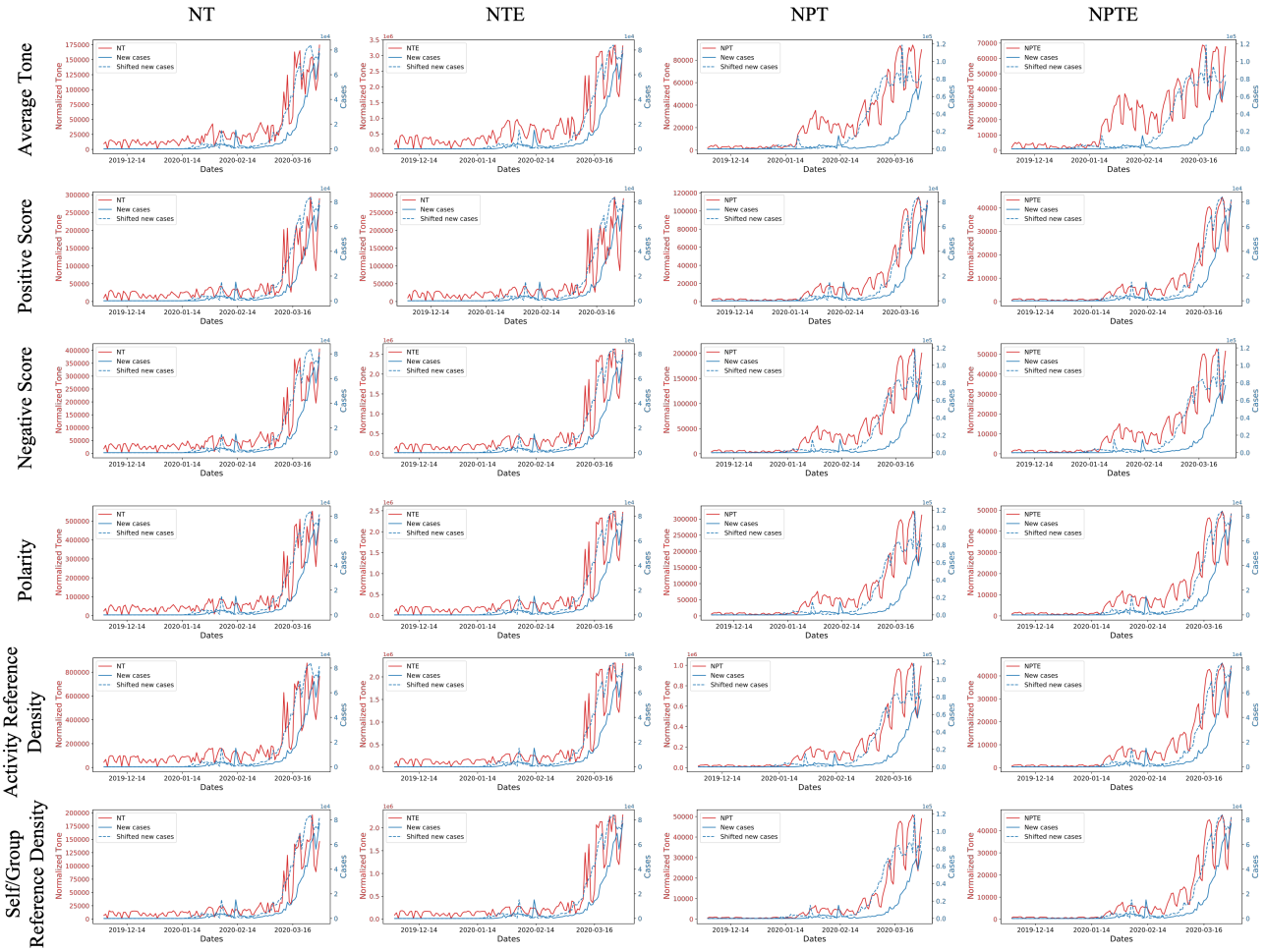


Figure 9. The normalized metrics for each emotional dimension. The red line indicates computed metrics (NT, NTE, NPT, and $NPTE$), and the blue lines indicate the newly infected cases of COVID-19 (solid) and the shifted graph of newly infected cases (dashed).