Generating Similarity Map in COVID-19 Transmission Dynamics with Topological Autoencoder

Pitoyo Hartono *

April 6, 2020

1   Introduction

At the end of 2019 the world has seen the initial outbreak of COVID-19, the disease caused by the SARS-CoV2 virus, in China. On March 22 2020 this disease has been declared as a pandemic by the World Health Organization with close to 300000 cases as of that day.

Artificial Intelligence communities have been starting to work to tackle many aspects of COVID-19 crisis as nicely reviewed in [1]. Data from this outbreak have already been compiled and made available for research communities [2].

In this short note, the author proposes the utilization of a topological neural network to map the globally dynamics of COVID-19 spread. With so many countries affected by this crisis, and each country has different geographical, demographic and strategic aspects, hence, it is very difficult to obtain clear global view on how those differences are correlated with the failure or success of mitigating this crisis.

The objective of this study is to offer a visualization method based on the time series data for COVID-19 patients numbers in many countries. While this visualization method does not directly offer insights for the means of crisis mitigation, it will offer an intuitive information on the similarities and differences between many countries in the transmission dynamics of this pandemic that will be useful for further analysis and prediction.

2   Topological Autoencoder

The outline of Topological Autoencoder (TA) is shown in Fig. 1. TA is a hierarchical neural network having a low dimensional topological hidden layer that is a simplified version of Soft-supervised Topological Autoencoder (STA) proposed in [3]. Originally, STA can be trained in supervised, unsupervised or the mix of both, and produces a two-dimensional topological map that allows human user to visualize the concept of the data. TA is a simplified version of STA in that it is only trained as an autoencoder in a unsupervised manner and produces a visible map that captures the inherent characteristics of the training data.

In this preliminary study, TA is trained with time series of the numbers of COVID-19 patients in more than 240 countries or regions in the world that have been daily compiled by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [4] and the data are made available in a GitHub repository [5].

The hidden layer of TA is a topological map similar to Kohonen's Self Organizing Maps (SOM) [6], [7], where hidden neurons are arranged in two-dimensional grid, and each neuron is associated with a reference vector that has the same dimensionality with the input.

The dynamics of TA is briefly explained as follows.

For a time series containing the daily numbers of COVID-19 in a certain country, $\mathbf{X} \in \mathbb{R}^d$, TA selects the best matching unit, win among all the reference vectors associated with the hidden units of STA as in Eq. 1, where $\mathbf{W}_j \in \mathbb{R}^d$ is the reference vector associated with the j-th hidden unit. Here, d is the number of the days from the beginning of data collection on January 22 2020.

The output of the j-th hidden neuron, H_j is shown in

---

*School of Engineering, Chukyo University, Nagoya, Japan

Eq. 2, where $\sigma(j, win, t)$ is the neighborhood function defined in Eq.3. Here, $\sigma_0 > \sigma_\infty > 0$ are the initial and final values of the annealing term, t is the current epoch, while t_∞ is the termination epoch.

$$win = \arg\min_j \|\mathbf{X} - \mathbf{W}_j\| \qquad (1)$$

$$H_j = \sigma(j, win, t)e^{-\frac{\|\mathbf{x} - \mathbf{w}_j\|^2}{\sigma^2}} \qquad (2)$$

$$\sigma(j, win, t) = exp(-dist(win, i)/S(t)) \qquad (3)$$
$$S(t) = \sigma_\infty + \frac{1}{2}(\sigma_0 - \sigma_\infty)(1 + cos\frac{\pi t}{t_\infty})$$

The values of the k-th output neuron, O_k, is defined in Eq. 4 where $f(x) = \frac{1}{1+e^{-x}}$

$$O_k = f((\mathbf{V}_k)^t \mathbf{H})$$

Here, $\mathbf{V}_k$ denotes the weight vector leading from the hidden layer to the k-th output neuron, $\mathbf{H} = (H_1, H_2, \cdots, H_{N_{hid}})^t$ is the hidden layer output vector in which N_{hid} is the number of hidden neurons.

The loss function is defined in Eq. 4.

$$L = \frac{1}{2}\sum_k (O_k - X_k)^2 \qquad (4)$$

TA is then trained by minimizing the loss function with the standard stochastic gradient descent. The detail for the loss function derivations is explained in detail in in [3], [8].

The modifications of the reference vectors in the hidden layer of TA cause the formation of two-dimensional topological map, in which countries with similar transmission dynamics are assigned adjacently, while countries with dissimilar dynamics are separated by large space on the map.

Utilizing this similarity map, we can gain intuitive understanding on global trend in COVID-19 transmission. The insights gained from visually analyzing this map are potentially helpful in designing strategy for controlling and thus mitigating the spread of this disease.

3 Topological Maps for COVID-19 Transmission

The following figures show a series of topological maps that reflect the dynamics of patients spread for around 240 countries or areas based on patients' numbers collected from January 22 2020 up to the date specified in the respective map. Each country is represented with a number according to the data numbering in John Hopkins University Repository. For visual clarity, the names of some countries or areas are shown on the map.

Figure 2 shows the topological maps for patients spread up to February 20 2020, the early stage when COVID-19 patients were detected outside China. From this map, it can be observed that many countries occupy the upper left area on the map. The overlapping are due to sharing of similar dynamics in the COVID-19 patients among many countries, in that most of them have identified very few patients at this time. At this time, Hubei, the epicenter for this epidemy, was already an outlier on the map, as the number of patients in Hubei was in different magnitude compared to other areas. It can be also observed here that USA, France, South Korea, and Asian countries as Malaysia, South Korea, Singapore, Thailand, Taiwan and Hong Kong were positioned adjacently, due to their similar dynamics, while many China's provinces as Zhejiang, Shanghai, Liaoning and naturally Hubei were outliers due to their early outbreaks.

Figure 3 shows the map up to March 6. The upper left of the map was still crowded, but many countries can be seen to breakout from this overlapping area. This is due to many countries were beginning to develop different strategies resulting in different patients transmission dynamics. It is notable that Italy and Iran moved considerably from the crowded cluster due to the beginning of outbreaks in these two countries, while USA still shared the space, and thus has similar dynamics with Macau and Taiwan.

Figure 4 shows the map up to March 25, a few days after the World Health Organization (WHO) has declared COVID-19 as a pandemic. The left corner became less crowded as more and more countries developed their unique spread dynamics. Noticeably, South Korea moved away, due to rapid surge of virus transmissions in this period. Outbreaks occurred in USA, Spain and Italy as re-
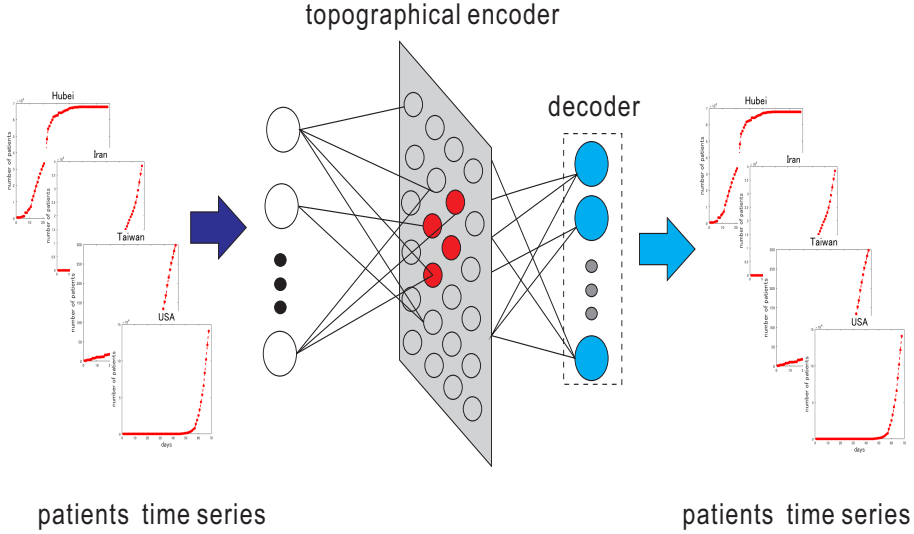
Figure 1: Topological Autoencoder

flected by their positions on the map. Hubei, the original epicenter for this disease, many of Chinese's provinces and South Korea kept their unique positions on the map because of their unique dynamics that saw early outbreak but than stabilized.

Figure 5 shows the map up to March 30, that did not significantly change from the previous map. This is because no noticeable new outbreaks, and hence many countries have stabilized their policies that results in the stabilization of their spread dynamics. The noticeable change is the split of Germany and Spain that previously shared a same dynamics.

Detailed analysis for this final map was executed to give better understandings on the topological characteristics of these maps.

Figure 6 shows the actual time series for some countries divided based on the region they occupied on the map in Fig. 5. From the left area, as shown in Fig. 6a, Beijing, Zhejiang and Vatican had dissimilar dynamics that were reflected by their large distances on the map. Beijing and Zhejiang have both flat out their spread but had different dynamics and patient numbers, while Vatican had very few patients up to March 30. In the middle of the map Japan, Netherlands and Switzerland can be observed each with considerably different dynamics as shown in Fig. 6b.

From the right side of the map, Hubei, USA and Italy as shown in Fig. 6c. It is obvious that Hubei had stabilized while patients numbers were growing steeply in USA and Italy. Figure 7 shows the internal representations, each one is the reference vector for the winning neuron associated with each country. Figure 7a shows the representation for Beijing's, Zhejiang's and Vatican's dynamics, Fig. 7b shows the representation for Japan's, Netherlands' and Switzerland's dynamics, while Fig. 6c shows those of Hubei, USA and Italy. Here, it is important to notice that TA represents different dynamics with different internal representations that are important to recreate the original dynamics in the output layer of TA. Here, it is important to notice that some similar, although different dynamics may be represented by a single representation. The representation capture the most important character of those dynamics, and thus becomes a building block for them.

To further explain the internal representation characteristics, the dynamics of South Korea and UK are shown in Fig. 8b. These two countries are deliberately chosen because of their contrasting transmission dynamics that are reflected by their largest distance on the map. South Korea are known for their success in containing the spread of this disease after an early outbreak, while UK, on March 30, was still struggling. Their internal representations are
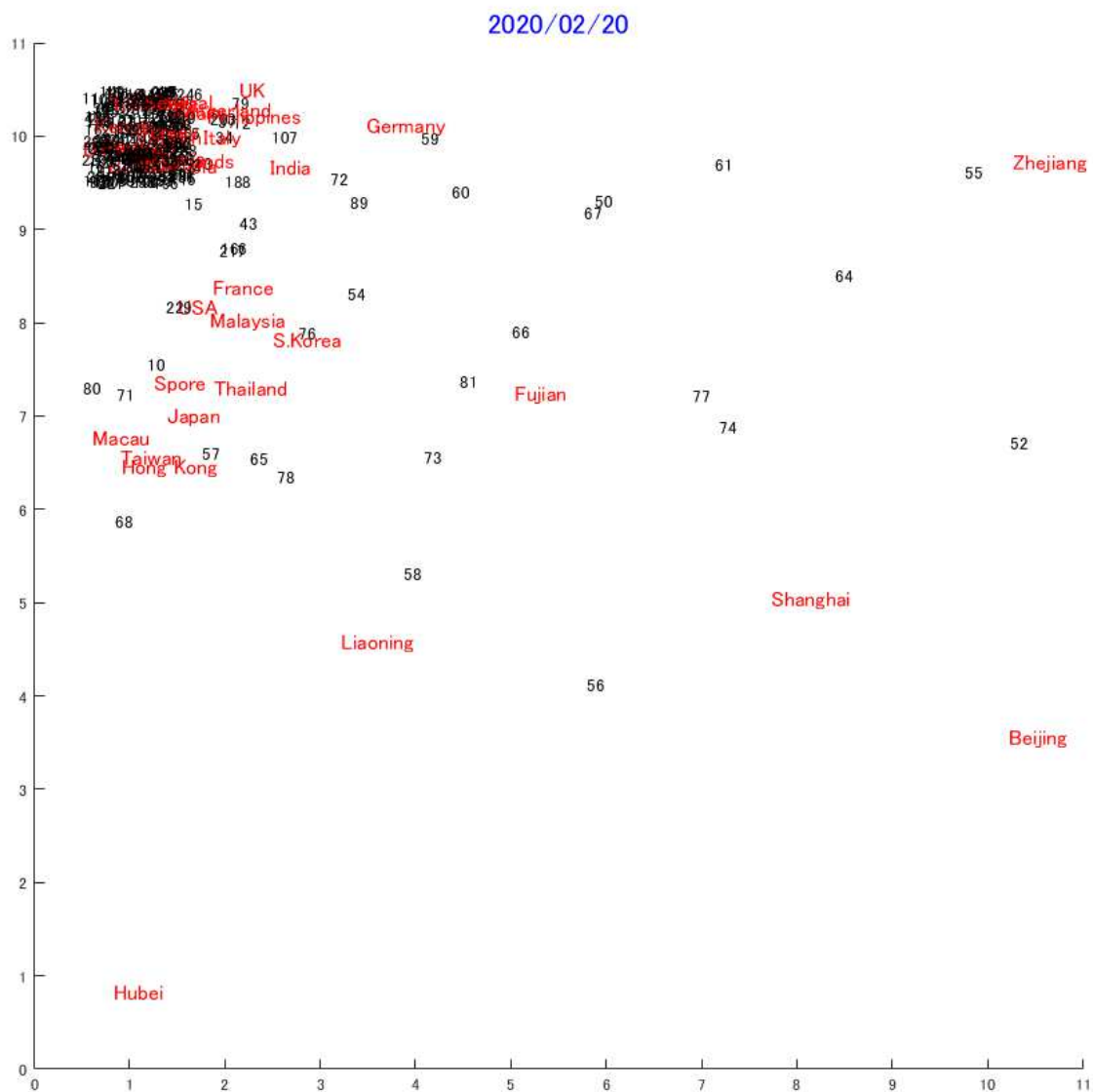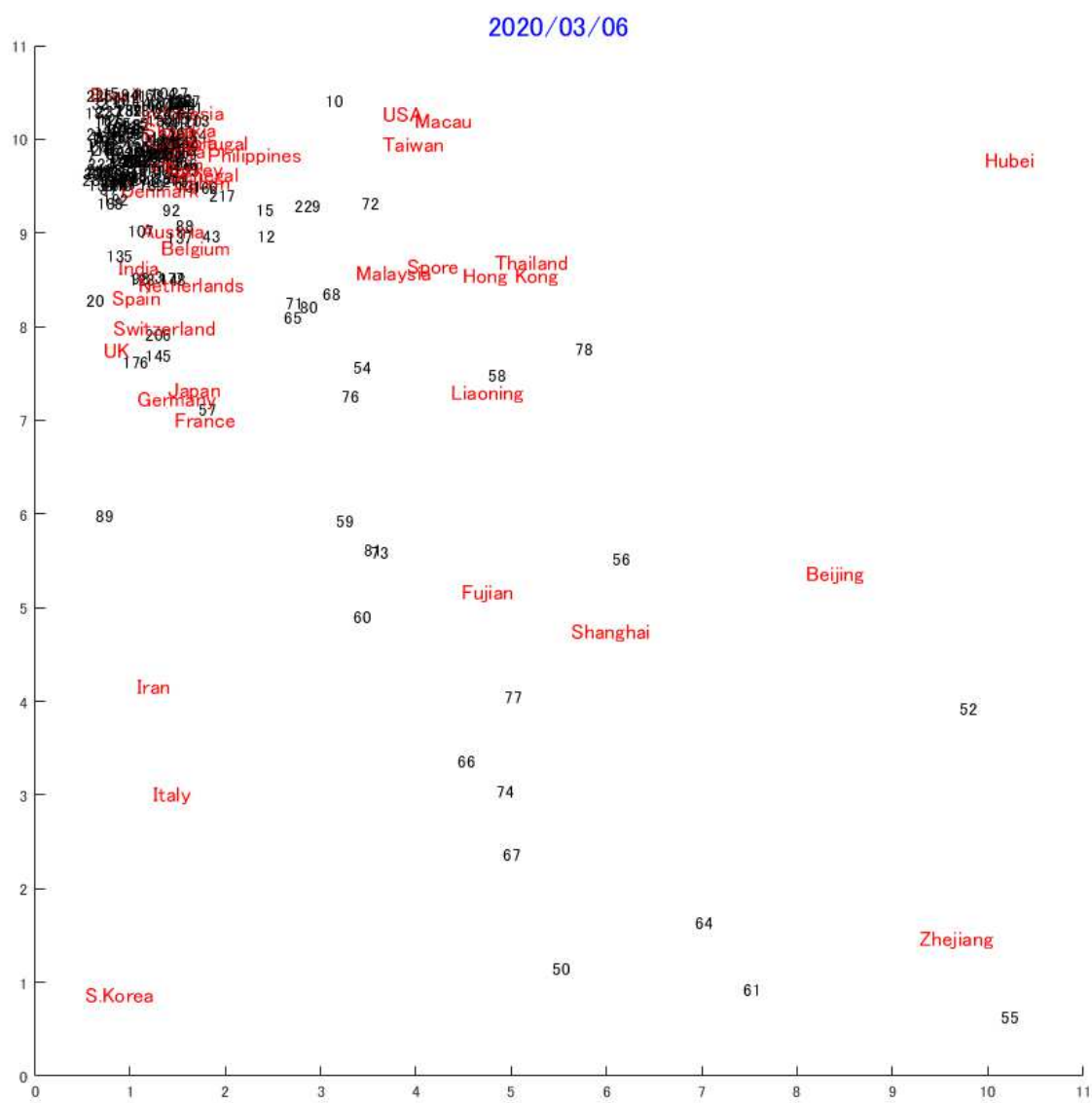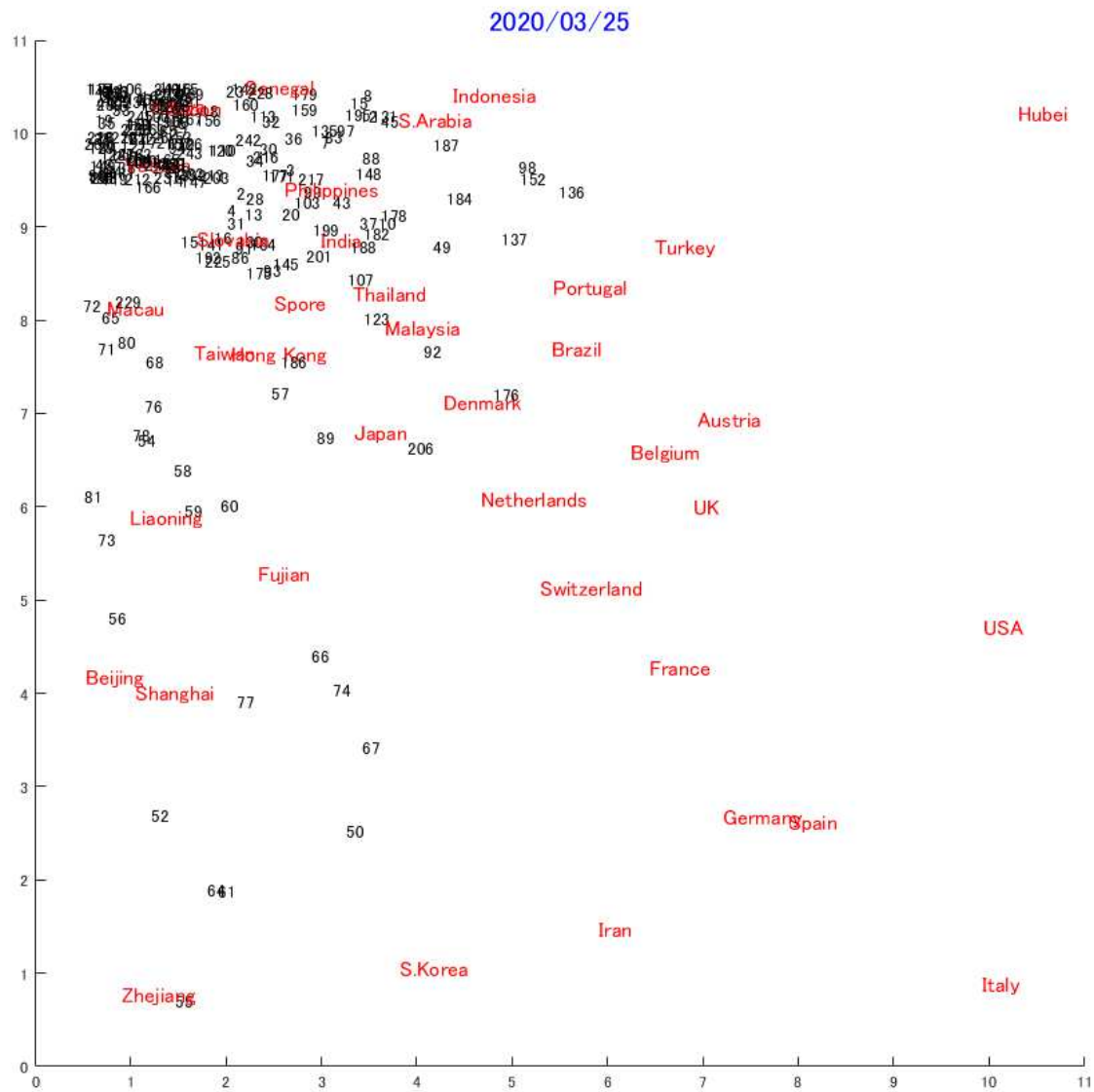
3

Figure 2: Map Feb. 20 2020

Figure 3: Map March 6 2020

5

Figure 4: Map March 25 2020

2020/03/30

Hubei

Indonesia
Philippines
S.Arabia
Macau
Slovakia

Denmark
Spore Malaysia Brazil
Hong Kong Taiwan Portugal
Thailand Turkey
Japan
Belgium
Liaoning
Netherlands
Fujian Austria USA
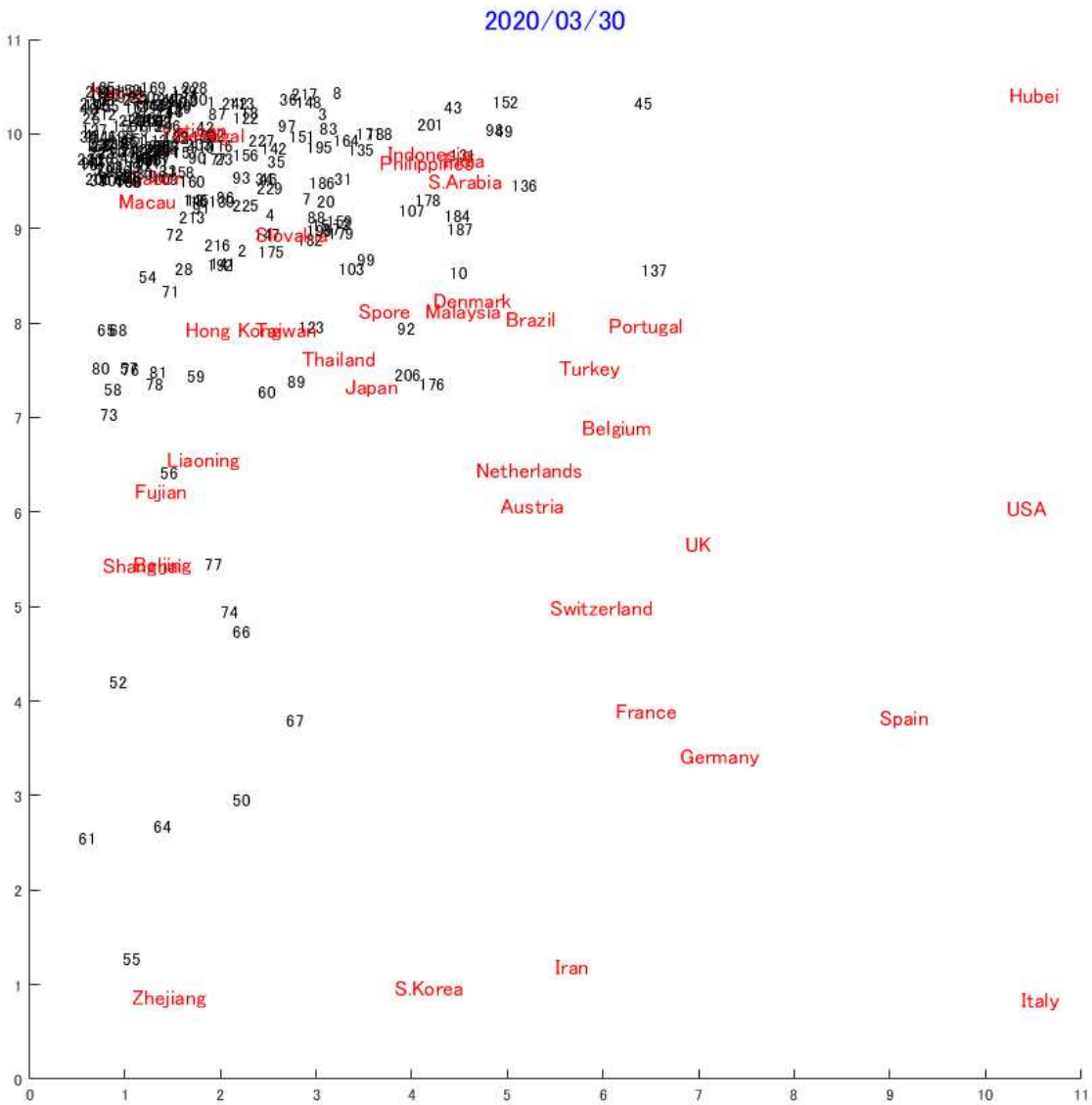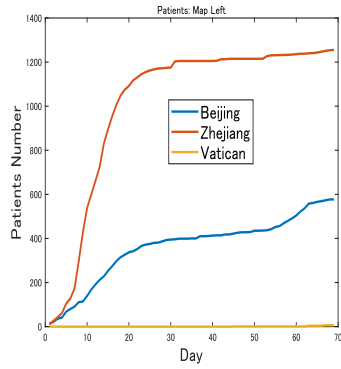Shanghai Beijing UK
Switzerland
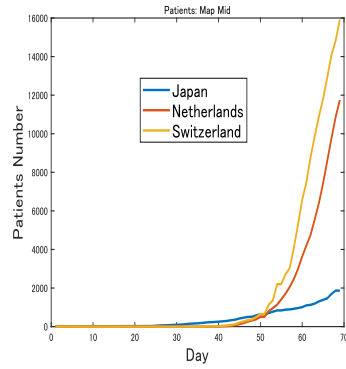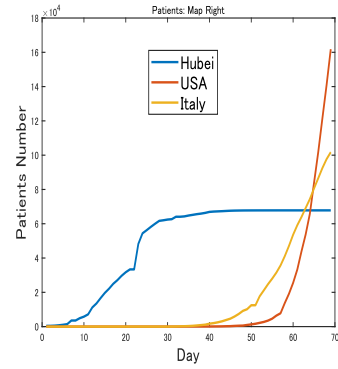France Spain
Germany
Iran
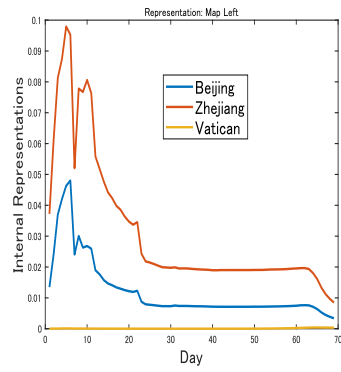Zhejiang S.Korea Italy

Figure 5: Map March 30 2020
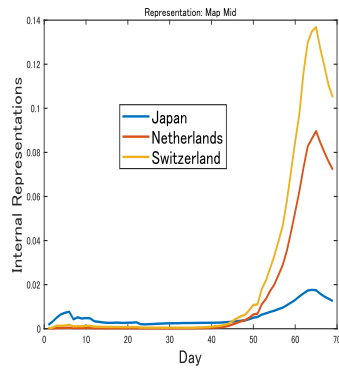
7

(a) Map Left: Patients   (b) Map Mid: Patients   (c) Map Right: Patients
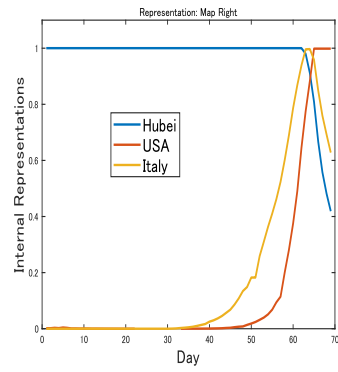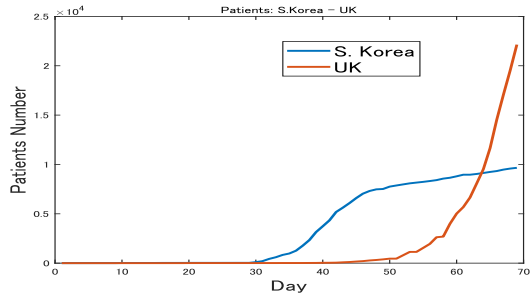
Figure 6: Patients Spread

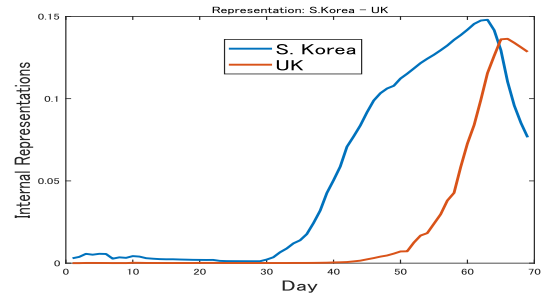(a) Map Left   (b) Map Mid   (c) Map Right
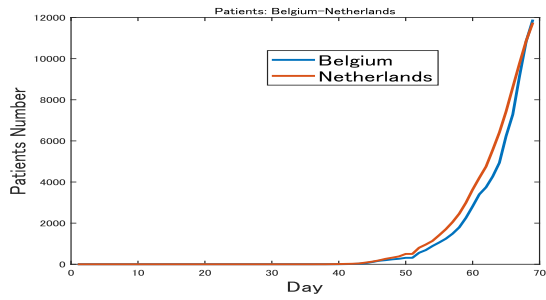
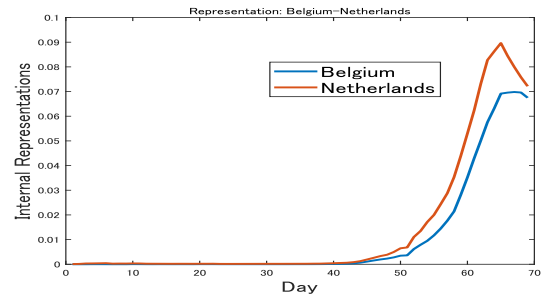Figure 7: Internal Representations

(a) Patients: S.Korea-UK

(b) Representation: S.Korea-UK

Figure 8: S. Korea vs UK

(a) Patients: Belgium-Netherlands

(b) Representation: Belgium-Netherlands

Figure 9: Belgium vs Netherlands

shown in Fig. 8b, which clearly show the difference between these two countries. Here a clear example for the reflection of the differences in dynamics into the internal representations is shown.

Oppositely, Fig. 9a shows the similar dynamics for Belgium and Netherlands that naturally are placed closely on the map in Fig. 5. The similarity in the disease spread dynamics is reflected in the similar representations in Fig. 9b and their positions on the map as well.

Through these preliminary examples, it can be empirically shown that the hidden representation of TA is able to capture the topological relations in different countries dynamics and thus can be utilized as an analytical tool, for intuitively comparing the dynamics of many countries and potentially analyzing the correlation between those dynamics and strategies to mitigate this crisis.

4 Conclusion

In this unrefereed note, the author proposes the utilization of topological neural network for visualizing the global trend for the COVID-19 transmission. The initial experiments indicate that we can gain intuitive insights on the similarities and differences in the disease transmission among many countries. By visualizing the maps in time, we can also learn the shifts in each countries situations, which mirror their strategies. At this moment, this topological map is not directly useful for evaluating the efficiency of strategies executed by many countries to mitigate the crisis. However, the author believes that the intuitive visual understanding gained from these topological maps will be useful in comparing strategies among many countries, and thus for designing optimal strategies for mitigating this crisis. The direct usage of this map, is to cluster countries having similar dynamics, and train a time series predictor, for example Long Short-term Memory (LSTM) [9], for further predicting the future numbers of patients in those countries.

References

[1] J. Bullock, Alexandra, Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, "Mapping the landscape of artificial intelligence applications against covid-19," 2020.

[2] B. Xu, B. Gutierrez, S. Mekaru, *et al.*, "Epidemiological data from the covid-19 outbreak, real-time case information," *Scientific Data*, vol. 7, no. 106, 2020.

[3] P. Hartono, "Mixing autoencoder with classifier: conceptual data visualization," *ArXiv e-prints*, 2019.

[4] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet Infectious Diseases*, 2020.

[5] "2019 novel coronavirus covid-19 (2019-ncov) data repository by johns hopkins csse." https://github.com/CSSEGISandData/COVID-19, 2020.

[6] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[7] T. Kohonen, "Essential of self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013.

[8] P. Hartono, P. Hollensen, and T. Trappenberg, "Learning-regulated context relevant topographical map," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2323–2335, 2015.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.