A Network-Based Approach to Predict New Affected Regions and the Spread Evolution of Covid-19

Tiago Colliri[a,*], Liang Zhao[b]

[a] *Dept. of Computer Science, Institute of Mathematics and Computer Science-USP, Sao Carlos, Brazil*
[b] *Dept. of Computing and Mathematics, Faculty of Philosophy, Science, and Letters-USP, Ribeirao Preto, Brazil*

Abstract

Given the most recent events involving the fast spreading of Covid-19, policy makers around the world are being challenged with the difficult task of developing efficient strategies to contain the dissemination of the disease among the populations, sometimes by taking severe measures that restrict the local activities both socially and economically. Within this context, models which can help on predicting the next regions to be affected by the disease and also its spread evolution in a specific region would surely help the authorities on their planning. However, the current prediction attempts in this sense are usually either for an unspecified time-range or by utilizing distinct models, focusing on a specific region. In this paper, we introduce two different network-based models by making use of preliminary available data regarding the spread of Covid-19. The first model predicts the new regions to be affected by the disease within a certain time range. It starts by mapping each region as a node in a network, and the edges are generated according to their proximity in terms of geographic coordinates. Afterwards, we apply link prediction techniques for generating the predictions within a predetermined number of days. The obtained experimental results on this task achieved an average accuracy of 90%, when predicting the next 10 regions to be affected within the next 21 days. The second model proposed in this work predicts the evolution of time series through temporal networks. In this case, each node represents a time series, and the edges are created according to the similarity of their variations at each time step. The results obtained by applying this model on time series concerning the spread evolution of Covid-19 on different world regions are promising, with the predictions being consistent with later real spread evolution data released for these same regions.

Keywords: complex networks, machine learning, classification, time series prediction, temporal networks, Covid-19

1. Introduction

Since the first human populations started to live in groups, the epidemic diseases have been one of the greatest problems faced by humanity. In the modern era, this problem has been aggravated by two factors combined: (1) the fact that human beings have been living more and more in concentrated urban spaces, and (2) these great concentrations of people, by their turn, are increasingly more interconnected through faster and more efficient worldwide transportation routes [1]. The most recent example regarding this problem involves the Covid-19, which was officially characterized as a pandemic only around four months after its first cases were reported, in China. The fast dissemination of this disease can be particularly challenging for policy makers around the world, which have been facing the difficult task of developing efficient strategies to control the spread of the disease among the populations, sometimes by taking severe measures that restrict both socially and economically the local activities. Within this context, models which can help on

*Corresponding author
Email addresses: tcolliri@usp.br (Tiago Colliri), zhao@usp.br (Liang Zhao)

predicting the next regions to be affected by this disease and also its spread evolution in a specific region would surely help the authorities on their resources planning. However, the current prediction attempts in this sense are usually either for an unspecified time-range [2] or by utilizing distinct models, focusing on a specific region [3,4,5,6].

Networks (or *graphs*) are powerful modeling tools for exploring a dataset in terms of the relations between the data instances, both in a static or in a dynamic way. The term *complex network* refers to a graph consisting of a large number of *nodes* (or vertices) joined by *links* (or edges), with a non-trivial topology [7]. Some examples of complex networks include the internet [8], biological neural networks [9], social networks among individuals [10], blood distribution networks [11] and power grid distribution networks [12]. There are also several network-based models designed to perform *machine learning* tasks, such as *clustering* [13], *classification* [14,15] and *regression* [16]. Mathematically, a network can be defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of tuples representing the edges between pairs of nodes $(i, j) : i, j \in \mathcal{V}$. A *temporal network* is a specific type of *multilayer network* or *multiplex* [17], in the form of $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{D})$, in which the additional dimension $\mathcal{D}$ contains an ordered set of temporal indices that represents time [18]. Among the phenomena which have already been modeled through temporal networks are person-to-person communication [19], brain connectivity [20], fires events in the Amazon [21], economic trade and social networks [22], political parties [23,24] and corruption scandals [25].

Although complex networks is a relative new field of study, there are already some well-known network-based models developed specifically to approach problems regarding the spread of epidemic diseases. These models do not necessarily need to make use of very advanced mathematical calculus since, oftentimes, simple models can help to further understand the transmission of infectious agents within human communities [26]. The SI, SIS and SIR models [27,28], for instance, allow one to estimate what would be the critical threshold, in terms of the percentage of infected individuals in a population, for an infectious disease to become endemic. These models also can help on determining which immunization strategies are expected to be more effective, according to the topological characteristics of the network formed by the individuals susceptible to the disease [29,30]. Other works [31,32,33] made use of real data from the *Mycoplasma pneumoniae* infection, HIV and avian influenza, respectively, to validate the spreading simulations on complex networks.. More recently, studies in the field of epidemic spreading have changed their focus to metapopulation approaches instead of individuals-based ones, where each node of the network may represent a set of individuals, such as a spatial region or a city, for example [34]. A very interesting survey on this topic was made by Costa et al. [35].

In this paper, we introduce two different network-based models by making use of preliminary available data regarding the spread of Covid-19. The first model predicts what are the new regions to be affected by the disease, within a certain time range in advance. It starts by mapping each world region as a node in a network, and the edges are generated according to each region's geographic coordinates. Afterwards, link prediction techniques are used to identify which are the next regions to be affected by the disease. The obtained results on this task achieved an average accuracy of 90%, when predicting the next 10 regions to be affected by Covid-19 within the next 21 days (3 weeks). The second model introduced in this work predicts the spread evolution of Covid-19 in a specific region through a temporal network. It starts by mapping the Covid-19 spread evolution data by world region to static networks, where each network represents the current spread variations by region in terms of their relations in the topological space, for each time step t. Each node in these networks represents a different region, and the edges are generated according to the similarities between the current variations in the total number of infected by the disease, for each time step t. Afterwards, these static networks are analyzed in the form of a temporal network, by considering them as an evolving network, in order to predict the evolution of the Covid-19 dissemination in a specific region. The obtained results in this task are promising, since the evolution data predicted by the model are consistent with real Covid-19 spreading data released after the predictions.

Regarding the organization of this paper, besides this introduction, we have, in section II, a description of the two models, showing how the networks are generated from the input data, and how the predictions and estimates of the model are made. In section III, we present the results obtained by applying the models to real preliminary data regarding the spread of Covid-19 around the world. In section IV, we conclude this study with some final remarks.

2. Methodology

The research methodology used in this study is summarized below.

2.1. Database

The database used in this study is built from the preliminary data made publicly available by Dong et al.[36]. This dataset comprises the daily evolution of Covid-19 confirmed cases on 250 different regions or countries, from 01-22-2020 until 03-30-2020, along with their respective latitude and longitude coordinates. We add in this dataset another 62 countries, which until 03-30-2020 did not have any Covid-19 confirmed cases, along with their respective geographic coordinates. Thus, we end up with a total of 312 different regions or countries in the dataset, with 250 of them being already affected by Covid-19 and another 62 of them not having any confirmed cases until the final date.

2.2. Prediction of new affected regions

The model starts by mapping each region in the dataset as a node in a network, and the edges are generated based on the geographic coordinates of each of them. Hence, we have that the overall distribution of the nodes in the network will somewhat resemble the distribution of the regions in a world map, with nearer regions also staying close to each other in the network. Afterwards, depending on the task to be accomplished, a different step is required.

Mathematically, we start with a set $X = \{x_1, x_2, x_3, ..., x_n\}$ where each element x_i represents a different geographic region in the database. Then, we will have that $X \mapsto \mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V} = \{1, ..., V\}$ is the set of vertices and $\mathcal{E}$ is the set of edges in the network $\mathcal{G}$. The connections between nodes are created using two traditional graph formation techniques in a combined form. The neighbors connected to a vertice i are given by:

$$N(x_i) = \begin{cases} \epsilon\text{-}radius(x_i), \text{ if } |\epsilon\text{-}radius(x_i)| > 0 \\ \text{NN}(x_i), \text{ otherwise} \end{cases}, \tag{1}$$

where ϵ is a predefined value measured in terms of geographic coordinates, ϵ-$radius(x_i)$ returns the set $\{x_j \in \mathcal{V} : d(x_i, x_j) \leq \epsilon$, and $\text{NN}(x_i)$ simply returns the nearest neighbor of vertice i, disregarding the geographic distance. Note that the NN technique is used solely in cases when there are no other node within the area delimited by ϵ radius, i.e., when $|\epsilon\text{-}radius(x_i)| = 0$.

The distance between two regions x_i and x_j in the database is calculated according to its Euclidian version, being yielded by:

$$d(x_i, x_j) = \sqrt{(x_i^{lat} - x_j^{lat})^2 + (x_i^{lon} - x_j^{lon})^2} \quad , \tag{2}$$

where x_i^{lat} stands for the latitude of x_i and x_i^{lon} stands for the longitude of x_i.

In order to forecast which regions are expected to be the next ones to be affected by Covid-19, the model makes use of link predictors. At the day t, predictions of new edges are performed in the network, using link prediction techniques. Next, we sort the link predictions by their respective weights, and take the 10 highest weighted predictions which, at the day t, have a currently infected region as a source node and a non-infected region as a target node, and consider their target nodes as being the next ones to be affected by the disease within the next n days. The accuracy on this task is then measured by comparing the predictions with the real data containing the daily evolution of confirmed cases per region, from the database, at the day $t + n$. The following link prediction techniques are used in this task:

- *Adamic Adar*[37]: based on the amount of shared links between two nodes,
- *Common Neighbours*: based on the number of neighbors shared by each pair of nodes,
- *Rooted PageRank*[38]: based on an algorithm developed for ranking the importance of website pages[39],
- *SimRank*[40]: based on the level of similarity of the structural context in which the nodes occur, and

- *Random*: used as the baseline, for comparison purposes.

All link predictors are implemented through the tool introduced by Guns[41].

Mathematically, the score of each prediction performed by the model is given by:

$$\begin{cases} 1, \text{ if } x^{\theta}_{i_{t+n}} > 0 \\ 0, \text{ otherwise} \end{cases}, \tag{3}$$

where x_i is the region of the target node for the link predicted at the day t, and $x^{\theta}_{i_{t+n}}$ yields the number of Covid-19 confirmed cases in x_i on the day $t+n$.

2.3. Prediction of the spread evolution in a region

For predicting the evolution of the spread of Covid-19 in a specific region, we introduce a model which generates predictions for a time series in a dataset based on the evolution of other time series in the same dataset. This model is intended to be applied on cases when the time series in the dataset present different initial dates, and hence also different lengths, and one wants to investigate the existence of possible correlations between them and, if that is the case, then to estimate future values for the time series based on these detected correlations.

In *machine learning* applied to time series, initially we have an input dataset comprising m instances and n time steps t, in the form of $\mathcal{X} = \{X_1, X_2, X_3, ..., X_m\}$, where each instance i consists of n elements, such that $X_i = (x_{i,t=0}, x_{i,t=1}, x_{i,t=2}, x_{i,t=3}, ..., x_{i,t=n})$, as in the following 2d array:

$$\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & ... & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & ... & x_{2,n} \\ ... & ... & ... & ... & ... \\ x_{m,1} & x_{m,2} & x_{m,3} & ... & x_{m,n} \end{bmatrix} \quad . \tag{4}$$

The model proposed in this work is indicated for cases when the initial dates (and consequently also the length) of the time series may presents different values in the dataset. Hence, the model starts by bringing all time series in $\mathcal{X}$ to a same starting point $t = 1$. Next, it calculates the variation $\delta X_{i,t}$ for each time series i at each time step t, which is yielded by:

$$\delta X_{i,t} = \frac{X_{i,t} - X_{i,t-1}}{X_{i,t-1}} \quad . \tag{5}$$

This provides us with a 2d variations array $\delta\mathcal{X}$, containing m instances and a maximum of $n-1$ columns or time steps, for each instance. Afterwards, each column $\delta\mathcal{X}_t$ in this array is mapped as a network G_t, where each node represents a time series and the edges between each pair of nodes are generated according to the similarities between their variations at the time t. The neighbors connected to each vertice i, in each network G_t, are given by:

$$N(i_t) = \epsilon\text{-}radius(i_t), \tag{6}$$

where $\epsilon\text{-}radius(i_t)$ yields a set of instances whose variations $\delta X_{i,t}$ are within the range $[-\epsilon, +\epsilon]$. Following, a community detection algorithm is ran in the networks, in order to group the time series with more similar variations, at each time step t. For this end, one can use, for instance, the *fast greedy* algorithm[42], which detects the community structure based on the greedy optimization of modularity measure, or also the *walktrap community detection* algorithm[43] which, roughly speaking, is based on the idea that short random walks tend to stay in the same community in the network. At this point, we will end up with a set of static networks $\mathcal{G} = \{G_{t=2}, G_{t=3}, G_{t=4}, ..., G_{t=n}\}$, with each of them representing the topological space emerged from the current relations between the variations δX at the time slice t. Hence, we can also say that this set forms the temporal network $\mathcal{G}$, and each element represents a different time slice t of the time series' temporal network $\mathcal{G}$.

There is also a possibility, at this step of the model, to deal with outliers in the dataset, by changing the variation values in each array δX_t through a transforming function f, according to:

$$f(\delta X_{i,t}) = \begin{cases} Q(\delta X_t, h_0), \text{ if } \delta X_{i,t} < Q(\delta X_t, h_0) \\ Q(\delta X_t, h_1), \text{ if } \delta X_{i,t} > Q(\delta X_t, h_1) \\ \delta X_{i,t}, \text{ otherwise,} \end{cases} \tag{7}$$

where $Q(A, n)$ yields the n-th quantile in the array A. This option can be used for long-tails distributions, by fitting each array δX_t into a range $[Q(\delta X_t, h_0), Q(\delta X_t, h_1)]$.

Following, a dictionary D_i is created, for each instance i in the dataset, in a list $\mathcal{D} = \{D_1, D_2, D_3, ...D_m\}$. The set of keys K in D_i are given by all instances j in the dataset which have shared the same community with i, at any time step t, i.e., in any of the networks in $\mathcal{G}$, and whose lengths are greater than the length of i. The set of values V in D_i, for any key j, are given by the respective number of times that instances i and j have shared the same community in $\mathcal{G}$. Mathematically, we have that a dictionary can be defined as:

$$D \subseteq \{(k, v) \mid k \in K \wedge v \in V \wedge \forall (q, w) \in D : k = q \rightarrow v = w\} \quad , \tag{8}$$

and the set of keys K and set of values V in D_i are yielded by:

$$D_i^K = \{j \mid j \in G_{t,i}^C\} \mid t \in [2, n] \wedge j \in [1, m] \wedge u(j) > u(i) \wedge j \neq i, \quad \text{and} \tag{9}$$

$$D_i^V = u(\{G_t \mid j \in G_{t,i}^C\}) \mid t \in [2, n] \wedge j \in [1, m] \wedge u(j) > u(i) \wedge j \neq i \quad , \tag{10}$$

where $u(A)$ returns the length of array A and $G_{t,i}^C$ provides a set with all instances that share the same community with i in the network G_t.

Finally, the predicted variation for a time series i, at the time step $t > 2$, is equal to the averaged variations of the time series which are in the keys of the dictionary D_i, weighted by their respective values in D_i. Thus, the predicted variation $\hat{\delta} X_{i,t}$ is given by:

$$\hat{\delta} X_{i,t} = \begin{cases} \frac{\sum_j D_i^V[k \rightarrow j] \ \delta X_{j,t}}{\sum_j D_i^V[k \rightarrow j]}, \forall j \in D_i^K \mid j \in G_t, \text{ if } \{j \mid j \in D_i^K \wedge j \in G_t\} \neq \emptyset \\ \emptyset, \text{ otherwise.} \end{cases} \tag{11}$$

Therefore, the model is able to predict variations for a time series i, on a time step t, only if at least one of the other time series j in the dataset, which the model identified as having a similar evolution to i, has a length equal or longer than t. Note that the model, hence, utilizes a form of regression analysis, in which it does not take into account the time series i – which is the dependent variable – prior evolution curve for making the predictions, and the independent variables (or predictors) in this case are the evolution curves from longer time series in the dataset identified as being similar to i. In this sense, it is worth highlighting the important role played by the ϵ radius threshold parameter in the model, used for generating the edges in the networks. Smaller values of ϵ make the predictions performed by the model more sensitive to local averages in the dataset while, conversely, higher values of ϵ result in the model considering broader averages when making the predictions. It is also worth noting, in Eq. 11, that by weighting the variations of time series similar to i by the number of times this similarity was identified by the model, prior to the time t of the prediction, we are here assuming that the time series i tends to continue presenting these same similarities in the future.

2.3.1. Model's demonstration through a simple example

In order to illustrate how the proposed time series prediction model works, we now present its application on a simple dataset, to be used as example. Let us consider a dataset $\mathcal{X}$, comprising five time series: X_1, X_2, X_3, X_4 and X_5, with different initial dates and different lengths, as shown in Table 1. Since the model makes use of data from the longer time series in the dataset, in order to perform the predictions, then in this case it will attempt to predict future values only for X_3, X_4 and X_5. One can observe, from

Table 1: Example Dataset

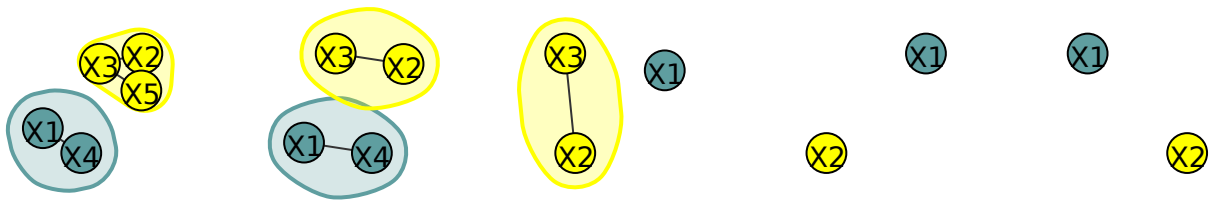
	X_1	X_2	X_3	X_4	X_5
1/1/2020	10	2	-	-	-
1/2/2020	15	4	-	-	-
1/3/2020	20	8	3	-	-
1/4/2020	25	16	6	6	-
1/5/2020	30	32	12	9	1
1/6/2020	35	64	24	12	2

Table 2: Daily Variations (%) for the Example Dataset

	X_1	X_2	X_3	X_4	X_5
1st day	50.0	100.0	100.0	50.0	100.0
2nd day	33.3	100.0	100.0	33.3	
3rd day	25.0	100.0	100.0		
4th day	20.0	100.0			
5th day	16.7	100.0			

Table 1, that X_1 and X_4 follow an arithmetic progression, while X_2 and X_3 follow a geometric progression, and X_5 – which is the most recent one, having only 2 observations so far – could either follow an arithmetic or a geometric progression in the future. However, let us suppose that, in the current context, we are not aware of these evolution patterns for any of the time series in the dataset, and hence we expect the model to correctly detect these evolution patterns for us and, afterwards, to also estimate what will be the future values for X_3, X_4 and X_5 solely by taking into account the observations from the other series in the dataset.

The first step in the proposed model is to bring all time series to a same starting point $t = 1$, and then to calculate their daily variations, as it is shown in Table 2. Next, the model generates 5 networks, i.e., the maximum length of the series in $\mathcal{X}$ subtracted by 1, where each node represents a time series in $\mathcal{X}$ and the edges between them are created according to the similarities of their daily variations, on each time step t. For accomplishing this task, in this example, we make use of the nearest neighbor technique based on a radius $\epsilon = 10$. Then, the *fast greedy* community detection algorithm is ran in the networks (Figure 1). This results in a temporal network $\mathcal{G}$, formed by the set $\{G_t \mid t \in [1, 5]\}$. Note that, in this step, the model groups the time series with more similar variations, at each time step t, and, as t gets bigger, only the nodes from time series with longer lengths are left in $\mathcal{G}$. The edges evolution among the nodes of $\mathcal{G}$ is shown in Fig. 2, where each row represents one time series in the dataset and each column represents one time slice of the temporal network. The colors denote the community to which each node belongs, at each time slice. In this case, if a node has a white color, it means that this node is not in the temporal network at this time



(a) 1st day (b) 2nd day (c) 3rd day (d) 4th day (e) 5th day

Figure 1: Networks generated by the model for the time series in the dataset used as example, with their respective communities, denoted by the nodes colors. Each node represents a different time series, and the edges between each pair of nodes are created according to how similar is the variation between them, at each time step t. The set containing these five static networks $\{G_t \mid t \in [1, 5]\}$ forms the temporal network $\mathcal{G}$.

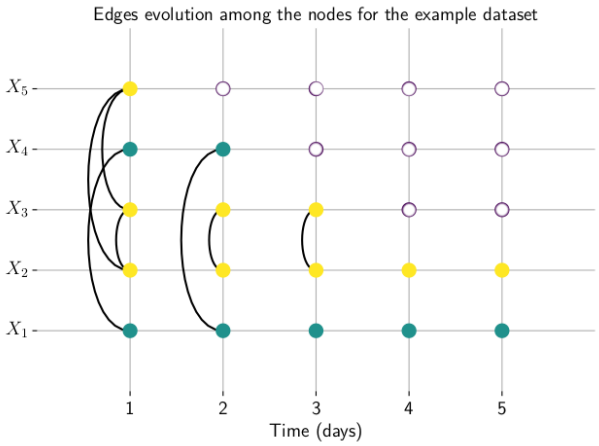
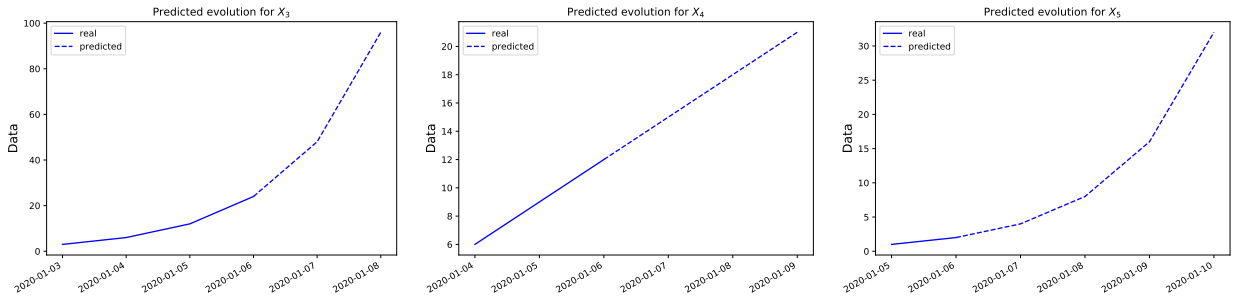


Figure 2: Time slices showing the edges evolution in the temporal network for the example dataset. Each row represents one time series in the dataset and each column represents one time slice of the temporal network. The colors denote the community to which each node belongs, at each time slice. In this case, if a node has a white color, it means that this node is not in the temporal network at this time slice.



(a) X_3 (b) X_4 (c) X_5

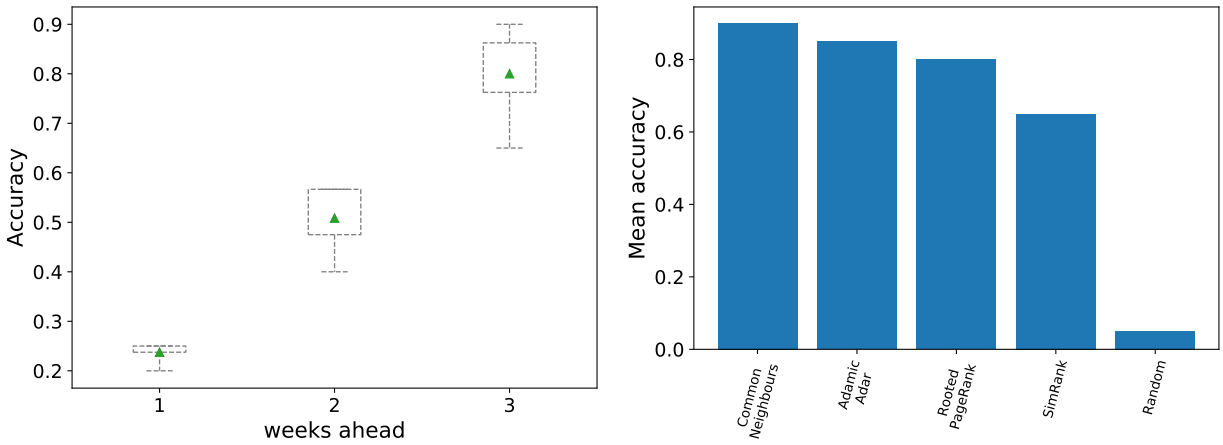
Figure 3: Predictions performed by the model for the time series (a) X_3, (b) X_4 and (c) X_5, from the dataset used as example. The blue line shows the series data provided in the dataset, while the blue dashed line indicates the predicted data.

slice.

Following, the model creates a dictionary for each time series i, with a set of keys containing the instances that have shared the same community with i whose lengths are greater than the length of i, and a set of values equal to the number of times they both have shared the same community. So, in this case, we have that the dictionaries for X_3, X_4 and X_5 are: $D_3 = \{X_2 : 3\}, D_4 = \{X_1 : 2\}$ and $D_5 = \{X_2 : 1, X_3 : 1\}$, respectively. Therefore, according to Eq. 11, the variation predicted for X_5 at the time step $t = 2$, is given by: $\frac{1\delta X_{2,2}+1\delta X_{3,2}}{1+1}$. In Fig. 3, we show all predictions made by the model for the example dataset. As one may observe, the model is capable to correctly detect and correlate the evolution patterns between X_1 and X_4 and between X_2 and X_3, and makes use of these detected correlations for making the predictions for X_3 and X_4. In the case of X_5, which could either evolve as an arithmetic progression or as a geometric one, the model ends up correlating it to X_2, and therefore the predictions for X_5 follow a geometric progression.

3. Covid-19 spreading prediction results

In this section, we present the obtained results when applying the network-based model to preliminary data available regarding the spread of Covid-19 around the world. We start by showing, in subsection 3.1,



(a) Predictions with 1, 2 and 3 weeks ahead

(b) Predictions with 3 weeks ahead

Figure 4: (a) Box plots of the accuracy achieved by the link prediction techniques Adamic Adar, Common Neighbours, Rooted PageRank and SimRank when predicting new regions to be affected by the disease, within the next 7 days (1 week ahead), 14 days (2 weeks ahead) and 21 days (3 weeks ahead). (b) Accuracy achieved by the same techniques when predicting new regions to be affected by the disease only within the next 21 days (3 weeks ahead). CommmonNeighbours predictor obtained the best performance in this task, with 0.9 accuracy. Note that the baseline Random model obtained only 0.05 accuracy in the same task. Period considered: from 01-22-2020 until 03-04-2020.

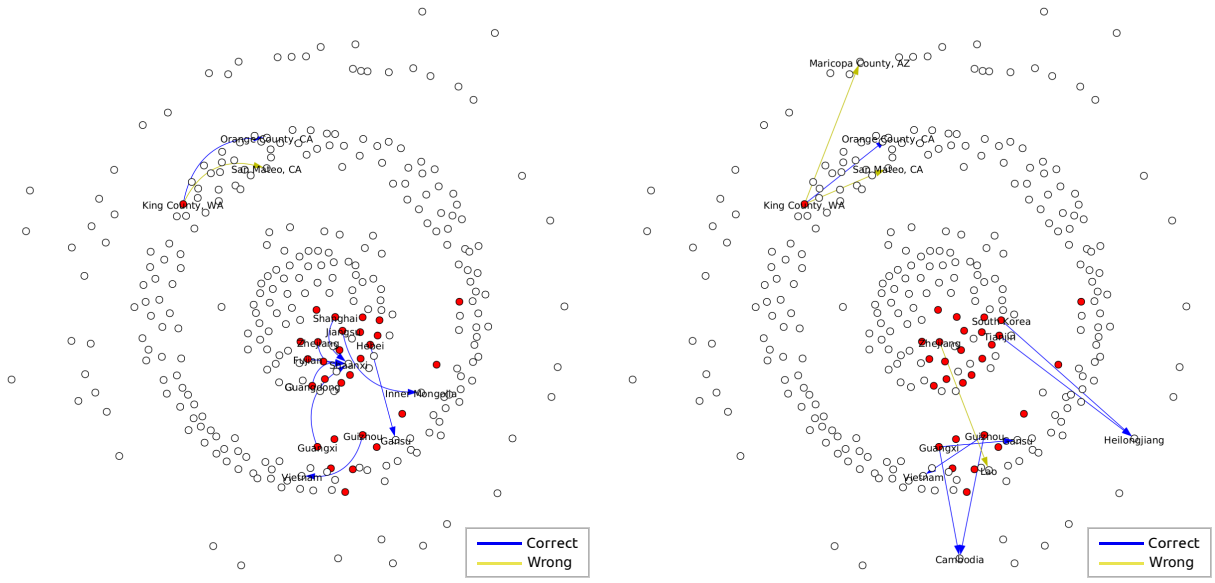
the results from the predictions of new regions to be affected by the disease. Then, in subsection 3.2, we show the obtained results when predicting the spread evolution of the disease in a specific region.

3.1. Prediction of new infected regions

The prediction of new regions to be affected by Covid-19 are made for the next 7 days, 14 days and 21 days, considering the first 42 days (or 6 weeks) in the database. The initial and final dates from the database used in this task are 01-22-2020 and 03-04-2020, respectively. On the final date, we have 156 regions already affected by the disease and the remaining 156 other regions in the dataset not affected by the disease. We perform 6 different predictions with 1 week ahead (7 days), 3 different predictions with 2 weeks ahead (14 days), and 2 different predictions with 3 weeks ahead (21 days). For building the network, we set the distance threshold parameter $\epsilon = 10$. The accuracy values achieved in each of these time intervals were averaged, and the results are summarized in Fig. 4(a). For predictions within the next 7 days, the accuracy is between 0.2 and 0.25. For predictions within the next 14 days it is between 0.4 and 0.57. For predictions within the next 21 days the accuracy is the highest, between 0.65 and 0.9. Hence, the obtained results on the Covid-19 preliminary data show that it is indeed possible to predict what will be the next regions to be affected by the disease within the next 3 weeks with a reasonable accuracy, with an average of 9 correct out of each 10 predictions made by Common Neighbours link predictor technique.

Still regarding Fig. 4(a), one can observe that long-term predictions present a higher precision than short-term ones. This can be explained by the fact that, on the long-term, more information regarding the spread of the disease is reflected in the network. Thus, as it is usually the case for machine learning models, the accuracy of the predictions tend to increase. Besides, we also have to take into consideration that, within this context, long-term data are more reliable than short-term ones, since the statistics of daily new confirmed cases in each region are subject to local determinant factors, such as limited testing capabilities, for example.

In Fig. 4(b), we show only the performances achieved on predictions made for the next 21 days (3 weeks), discriminated by each link prediction technique. Common Neighbours, Adamic Adar and Rooted PageRank predictors achieved an accuracy of 0.9, 0.85 and 0.8, respectively, on this task. Note that the



(a) Common Neighbours (9 correct, 1 wrong)

(b) Rooted PageRank (7 correct, 3 wrong)

Figure 5: Illustration of some link prediction cases performed by (a) Common Neighbours and (b) Rooted PageRank techniques in the network resulted from data regarding the infected regions on 01-22-2020 in able to predict new infected regions within the next 21 days (3 weeks). Currently infected regions are denoted by red nodes. The new regions to be infected are indicated by the edges targets. Blue arrows indicate a correct prediction and yellow arrows denote wrong predictions. For visibility purposes, only predicted edges and their respective source and target labels are shown in these figures.

baseline Random predictor achieved only 0.05 accuracy, which helps to demonstrate that the odds of making successful predictions in this task are low. Especially when we take into account the fact that there are 312 regions in total in the dataset, and the number of affected regions is 73 on 02-13-2020 and 156 on 03-04-2020, which are the respective dates used for checking the first and second predictions performed by the model.

In Fig. 5, it is possible to see some examples of predictions performed by the model in the network, on 01-22-2020. The edges showed in Figs. 5(a) and 5(b) are the links predicted by Common Neighbours and Rooted PageRank techniques, respectively. The next regions to be affected by the disease are the target nodes of each edge. Red nodes denote currently infected regions. We also displayed, in Table 3, the complete list of predictions made by each link prediction technique, on 01-22-2020 and 02-13-2020, along with their respective score. It is interesting to note, in this table, that all regions which appear more than once among the predictions made on a same day, by any of the techniques, are correct. The only exception for this rule is Myanmar which, although it appears 3 times among the predictions performed by SimRank on 01-22-2020, these predictions turned out to be wrong since, according to the database, no cases have been confirmed yet in this country until 03-04-2020.

3.2. Prediction of the spread evolution in a region

In order to apply the proposed time series prediction model on the Covid-19 database, we started by converting the daily confirmed cases variations to weekly variations. In this manner, given that there are 69 days in the database for this task (from 01-22-2020 until 03-30-2020), we ended up with a maximum of 9 weeks for the time series. Additionally, since in some of the regions in the database the times series actually started earlier than 01-22-2020, such as in many parts of China, in Japan and in the US, we decided to not make use of the time series from these regions in the application. This is because since we do not have the actual initial values for those time series, then we are not able to synchronize them with the other series in

Table 3: All link predictions made by Adamic Adar, Common Neighbours, Rooted PageRank and SimRank models, regarding the new regions to be affected by Covid-19 within the next 3 weeks, made on 01-22-2020 and 02-13-2020. The target column shows the predicted regions and the score values denote whether each prediction is correct (1) or wrong (0).

	01-22-2020			02-13-2020		
	source	target	score	source	target	score
Adamic Adar	King County, WA	San Mateo, CA	0	Italy	Macedonia	0
	King County, WA	Orange County, CA	1	Orange County, CA	Washington County, OR	1
	Jiangsu	Inner Mongolia	1	Orange County, CA	Umatilla, OR	1
	Guizhou	Vietnam	1	Los Angeles, CA	Umatilla, OR	1
	Shanghai	Shaanxi	1	Los Angeles, CA	Washington County, OR	1
	Guangxi	Shaanxi	1	King County, WA	San Mateo, CA	1
	Sichuan	Lao	0	San Diego County, CA	Umatilla, OR	1
	Ningxia	Inner Mongolia	1	San Diego County, CA	Washington County, OR	1
	Guangdong	Shaanxi	1	London, ON	Norfolk County, MA	1
	Zhejiang	Shaanxi	1	Italy	Poland	1
Common Neighbours	King County, WA	San Mateo, CA	0	Italy	Macedonia	0
	King County, WA	Orange County, CA	1	Orange County, CA	Washington County, OR	1
	Jiangsu	Inner Mongolia	1	Orange County, CA	Umatilla, OR	1
	Shanghai	Shaanxi	1	Los Angeles, CA	Umatilla, OR	1
	Guizhou	Vietnam	1	Los Angeles, CA	Washington County, OR	1
	Guangxi	Shaanxi	1	Italy	Poland	1
	Zhejiang	Shaanxi	1	King County, WA	San Mateo, CA	1
	Guangdong	Shaanxi	1	San Diego County, CA	Umatilla, OR	1
	Hebei	Gansu	1	San Diego County, CA	Washington County, OR	1
	Fujian	Shaanxi	1	Germany	Hungary	1
Rooted PageRank	King County, WA	Orange County, CA	1	Singapore	Indonesia	1
	King County, WA	San Mateo, CA	0	San Diego County, CA	Umatilla, OR	1
	Guizhou	Vietnam	1	Orange County, CA	Umatilla, OR	1
	Guangxi	Cambodia	1	Los Angeles, CA	Umatilla, OR	1
	Guizhou	Cambodia	1	San Diego County, CA	Washington County, OR	1
	Tianjin	Heilongjiang	1	Orange County, CA	Washington County, OR	1
	King County, WA	Maricopa County, AZ	0	Los Angeles, CA	Washington County, OR	1
	South Korea	Heilongjiang	1	Singapore	Brunei Darussalam	0
	Guangxi	Gansu	1	London, ON	Norfolk County, MA	1
	Zhejiang	Lao	0	Orange County, CA	Snohomish County, WA	1
Sim Rank	King County, WA	Orange County, CA	1	San Diego County, CA	Umatilla, OR	1
	Guizhou	Vietnam	1	Orange County, CA	Umatilla, OR	1
	King County, WA	San Mateo, CA	0	Los Angeles, CA	Umatilla, OR	1
	Tianjin	Heilongjiang	1	San Diego County, CA	Washington County, OR	1
	Sichuan	Bhutan	0	Orange County, CA	Washington County, OR	1
	Guizhou	Myanmar	0	Los Angeles, CA	Washington County, OR	1
	Hainan	Myanmar	0	Finland	Poland	1
	South Korea	Heilongjiang	1	San Diego County, CA	Snohomish County, WA	1
	Guangxi	Myanmar	0	Orange County, CA	Snohomish County, WA	1
	Sichuan	Lao	0	Tibet	Myanmar	0

the database, i.e., to bring all of them to a same starting point $t = 1$. Consequently, leaving them in the database would result in inaccurate variation predictions from the part of the model. We set the value of the ϵ radius threshold parameter, used for generating the edges in the networks, as $\epsilon = Q(\delta X_t, .08)$, where $Q(A, n)$ stands for the n-th quantile of the array A. For detecting communities in the network, we opted for the *walktrap community detection* algorithm. We also make use of a transformation function to deal with outliers in the dataset, as in the Eq. 7 of subsection 2.3, setting the parameters h_0 and h_1 to 0.25 and 0.75, respectively, i.e., the interquartile range.

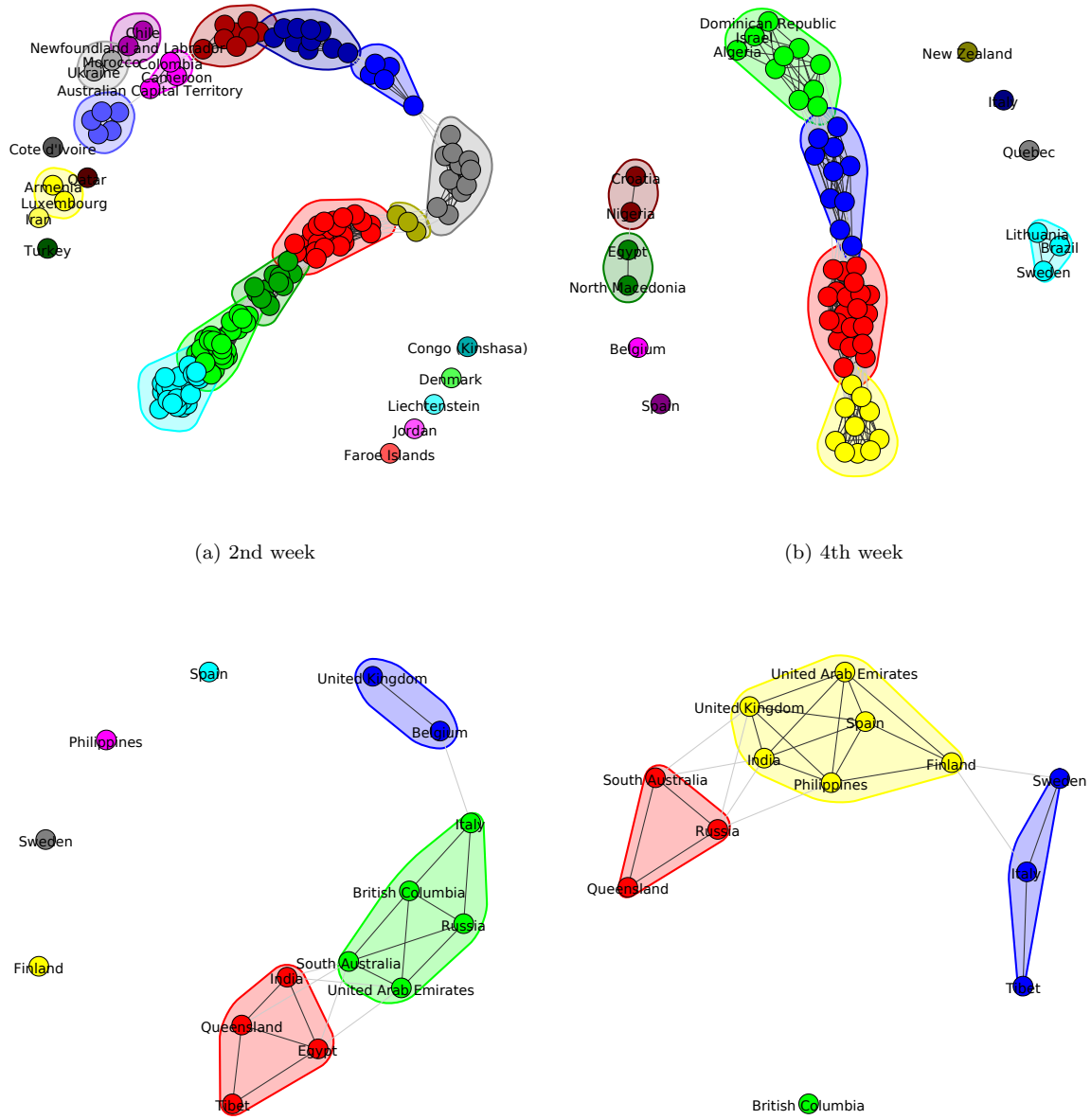
We start by showing, in Fig. 6, the static networks generated by the model for the dataset when $t = 2$ (2 weeks), $t = 4$ (4 weeks), $t = 6$ (6 weeks) and $t = 8$ (8 weeks). This figure is equivalent to the Fig. 1 from the example dataset. These static networks can be seen as time slices of a same temporal network $\mathcal{G}$. We can note, for instance, in Fig. 6(a), that some nodes – as Iran, Turkey, Qatar and Denmark – are not connected to any other node in the network. That means that, according to the model, these regions had a variation of confirmed cases of Covid-19 in the second week that were not similar to any other variations in the dataset. The same can be said for Italy, Spain, New Zealand, Quebec and Belgium, in the fourth week (Fig. 6(b)). As for the sixth week (Fig. 6(c)), the variation of confirmed new cases were similar for the United Kingdom and Belgium, and for India, Queensland, Tibet and Egypt, for example.

In Fig. 7, we present the evolution of confirmed new cases of Covid-19 predicted by the model for 20 different regions in the dataset. The blue line shows the real evolution prior to the predictions, the dashed blue line indicates the evolution predicted by the model, on a weekly basis, and the red line shows the actual daily evolution, released after the predictions. This figure is equivalent to Fig. 3 from the example dataset. Overall, we can say that the predictions made by the model are satisfactory, since it was also capable of correctly predict future shifts in the evolution curves, not only for the first week but for the second week as well, in some cases (as for Brazil and Channel Islands, for instance). It is also interesting to note that the predictions are consistent disregarding the existing differences between the number of confirmed cases, in absolute terms, in each region. The model is able to make valid predictions both for regions with a low number of confirmed cases, as in Sudan, with less than 20 cases, and for regions with a much higher number of confirmed cases, as in Belgium, Brazil, Poland and Romania, where the numbers are greater than 5000 cases.

The dictionaries generated by the model for making the predictions help us to understand how the future evolution is calculated, for each region. In the case of Panama, for example, its real data comprise two weeks in the dataset, so the model tried to predict the third and the fourth weeks for this region. Its dictionary has only two items, both with the same weight, in the form of $D_{Panama} = \{Alberta : 1, Iran : 1\}$. The model considered its variation in the first week similar to the variation of Iran (6800% and 6850%, respectively), and its variation in the second week similar to the variation of Alberta (400% and 404%, respectively). For the third week, the model considered both of these regions in the prediction, predicting a 240% growth in the total number of confirmed cases. For the fourth week, it predicted a 93% growth in cases, this time only taking the variation of Iran into account, since Alberta had no evolution data for the 4th week in the dataset. As for the case of Brazil, its real data comprise four weeks in the dataset, so the model made predictions for the fifth and sixth weeks. Its dictionary in the model is $D_{Brazil} = \{Queensland : 1, Oman : 1, Israel : 2, India : 1, Sweden : 1\}$, which means that the model identified a similar variation in the number of Covid-19 cases between Brazil and Israel two times, and one time between Brazil and the other regions in the dictionary. In Fig. 6(b), we can see that Brazil also shared the same community with Lithuania, in the network of the 4th week. However, since Lithuania's spreading data also comprise four weeks, it was not included in the dictionary for Brazil.

4. Final remarks

In this work, we made use of preliminary available data regarding the spread of Covid-19 in order to introduce two network-based models to analyze these data. The first model makes use of link prediction techniques to predict what are the new regions to be affected by the disease, by using link prediction techniques. The obtained results in this task show that it is possible to predict the next 10 regions to be affected, with 3 weeks in advance, with an average accuracy of 0.9. The second model introduced in this



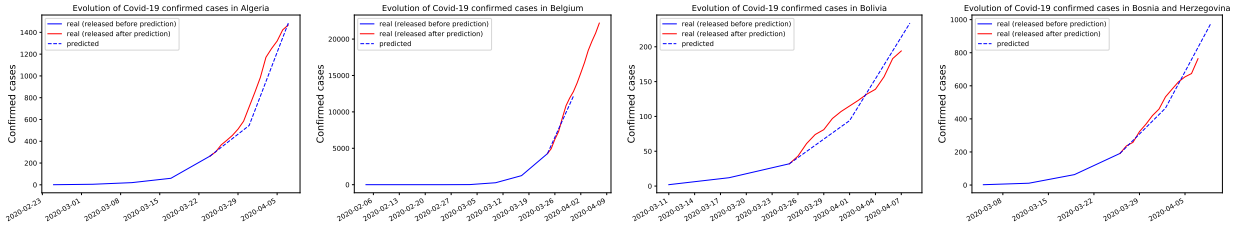
(a) 2nd week

(b) 4th week

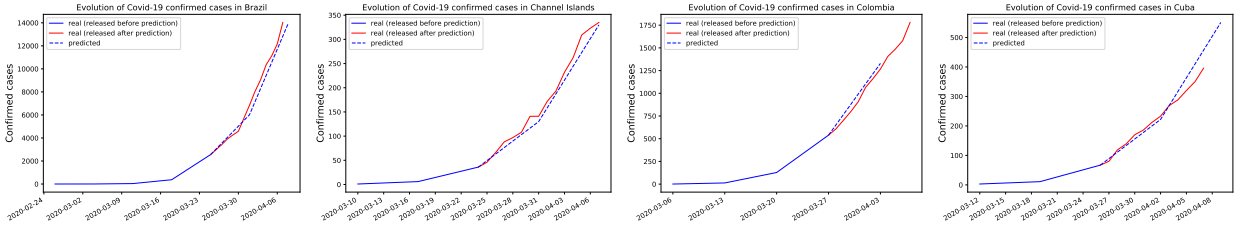
(c) 6th week

(d) 8th week

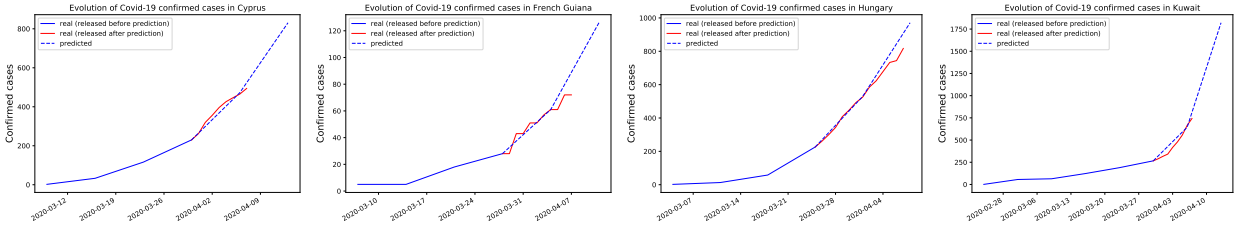
Figure 6: Networks generated by the model for the (a) 2nd, (b) 4th, (c) 6th and (c) 8th week, and their respective communities, since the first confirmed case of Covid-19 in each region. Each node represents a different region, and the edges between each pair of nodes are created according to how similar is the variation of new confirmed cases between them, in each week. These static networks can be seen as time slices of a same temporal network $\mathcal{G}$. For the sake of visibility, not all labels are shown in figures (a) and (b).



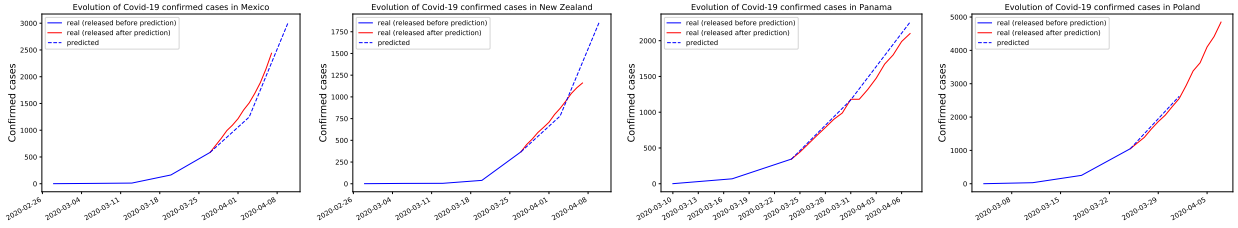
(a) Algeria (b) Belgium (c) Bolivia (d) Bosnia and Herzegovina



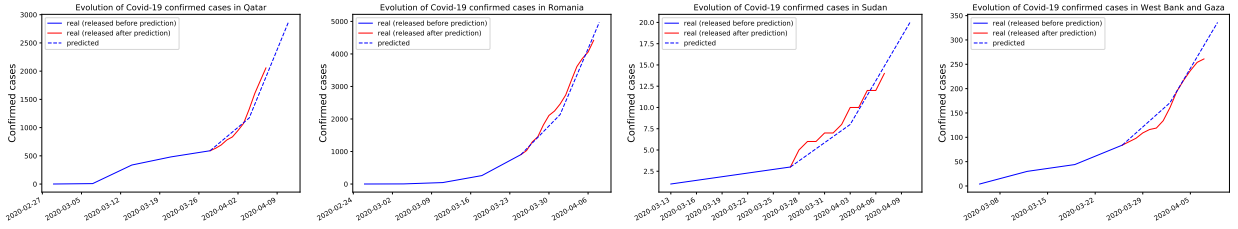
(e) Brazil (f) Channel Islands (g) Colombia (h) Cuba



(i) Cyprus (j) French Guiana (k) Hungary (l) Kuwait



(m) Mexico (n) New Zealand (o) Panama (p) Poland



(q) Qatar (r) Romania (s) Sudan (t) West Bank and Gaza

Figure 7: Examples of predictions performed by the model, on a weekly basis, for the evolution of confirmed new cases of Covid-19 in 20 regions. The blue line shows historical evolution data prior to the prediction, indicated by the blue dashed line. The red line shows daily evolution data released after the predictions.

work makes use of a temporal network and is intended to be applied on the prediction of time series in cases when the instances in the dataset present different initial dates and, henceforth, have different lengths. The application of this model on data regarding the Covid-19 spread evolution in different regions obtained satisfactory results, with the spread evolution predictions performed by the model being very consistent, when compared to the real evolution data released for these same regions after the predictions were made.

Regarding the spread evolution prediction task, we have to take into consideration that, given the nature of the data used in this study, the numbers of confirmed cases of Covid-19 in each region may not necessarily correspond to what would be the actual number of cases in these regions. This is because the accuracy of these numbers are subject to local factors, such as limited testing capabilities, for instance. This observation is valid both for when we compare different regions, as also for when we compare the testing capability within the same region on different time periods. There is also the fact that many people may not present any symptoms when infected by the disease. Moreover, many policy makers around the world have been taking severe measures during the period comprised in the database, by restricting the movement of people through lockdowns and quarantines, with the objective of flattening the local spreading curve of the disease. The effect of all these factors on the real data can make the task of predicting the evolution of these time series more challenging. Hence, the fact that the proposed model was still able to present satisfying predictions, despite the difficulties involved in this specific problem, corroborates to demonstrate that its rationale is valid and might as well be applied successfully to other types of time series datasets, from very diverse areas.

As future research, we believe it would be interesting to test the model based on link prediction techniques for detecting which areas should be the next to be affected by Covid-19 using more detailed data, within one country. In this way, one could predict, for example, which cities (or neighborhoods, depending on how detailed are the data) should be the next ones to be affected by the disease. As for the proposed time series prediction model, besides applying it to datasets from different areas for further testing its adaptability, as mentioned earlier, we also plan to explore other forms of analyzing its output data. One of the possibilities, in this sense, is to generate indexes for measuring how much the evolution of a time series is correlated to the global and local average evolution in the dataset, for instance.

Acknowledgements

This work is supported in part by the São Paulo State Research Foundation (FAPESP) under grant numbers 2015/50122-0 and 2013/07375-0, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and the Brazilian National Council for Scientific and Technological Development (CNPq) under grant number 303199/2019-9. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Y. N. Harari, Sapiens: A brief history of humankind, Random House, 2014.
[2] P. G. Walker, C. Whittaker, O. Watson, M. Baguelin, K. Ainslie, S. Bhatia, S. Bhatt, A. Boonyasiri, O. Boyd, L. Cattarino, et al., The global impact of Covid-19 and strategies for mitigation and suppression, Imperial College COVID-19 Response Team (2020).
[3] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al., Early dynamics of transmission and control of Covid-19: a mathematical modelling study, The Lancet Infectious Diseases (2020).
[4] A. Remuzzi, G. Remuzzi, Covid-19 and Italy: what next?, The Lancet (2020).
[5] Z. Liu, P. Magal, O. Seydi, G. Webb, Predicting the cumulative number of cases for the Covid-19 epidemic in China from early data, arXiv preprint arXiv:2002.12298 (2020).
[6] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan, G. Chowell, Real-time forecasts of the Covid-19 epidemic in China from February 5th to February 24th, 2020, Infectious Disease Modelling 5 (2020) 256–263.
[7] R. Albert, A. L. Barabási, Statistical mechanics of complex networks., Reviews of Modern Physics 74 (2002) 47–97.
[8] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology., ACM SIGCOMM Computer Communication Review 29 (1999).
[9] O. Sporns, Network analysis, complexity, and brain function., Complexity 8 (2002) 56–60.
[10] P. J. Carrington, J. Scott, S. Wasserman, Models and methods in social network analysis., Cambridge University Press, Cambridge, 2006.

[11] G. B. West, J. H. Brown, B. J. Enquist, A general model for the structure, and allometry of plant vascular systems., Nature 400 (2009) 125–126.
[12] R. Albert, I. Albert, G. L. Nakarado, Structural vulnerability of the north american power grid., Physical Review 69 (2004) 025103.
[13] T. C. Silva, L. Zhao, Stochastic competitive learning in complex networks, Neural Networks and Learning Systems, IEEE Transactions on 23 (2012) 385–398.
[14] T. C. Silva, L. Zhao, Network-based high level data classification., Neural Networks and Learning Systems, IEEE Transactions on 23 (2012) 954–970.
[15] T. Colliri, D. Ji, H. Pan, L. Zhao, A network-based high level data classification technique, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1–8.
[16] X. Gao, H. An, W. Fang, X. Huang, H. Li, W. Zhong, Y. Ding, Transmission of linear regression patterns between time series: From relationship in time series to complex networks, Physical Review E 90 (2014) 012818.
[17] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, A. Arenas, Mathematical formulation of multilayer networks, Physical Review X 3 (2013) 041022.
[18] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter, Multilayer networks, Journal of complex networks 2 (2014) 203–271.
[19] J. Tang, C. Mascolo, M. Musolesi, V. Latora, Exploiting temporal complex network metrics in mobile malware containment, in: 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, IEEE, pp. 1–9.
[20] W. H. Thompson, P. Brantefors, P. Fransson, From static to temporal network theory: Applications to functional brain connectivity, Network Neuroscience 1 (2017) 69–99.
[21] G. Xubo, Z. Qiusheng, D. A. Vega-Oliveros, A. Leandro, L. Zhao, Temporal network pattern identification by community modelling, Scientific Reports 10 (2020).
[22] D. Tamar, P. Kristijan, K. Ljupcho, Graphlets in multiplex networks, Scientific Reports (Nature Publisher Group) 10 (2020).
[23] T. Colliri, L. Zhao, Analyzing the bills-voting dynamics and predicting corruption-convictions among Brazilian congressmen through temporal networks, Scientific Reports 9 (2019) 1–11.
[24] S. Aref, Z. Neal, Detecting coalitions by optimally partitioning signed networks of political collaboration, Scientific Reports 10 (2020) 1–10.
[25] I. Luna-Pla, J. R. Nicolás-Carlock, Corruption and complexity: a scientific framework for the analysis of corruption networks, Applied Network Science 5 (2020) 1–18.
[26] R. M. Anderson, B. Anderson, R. M. May, Infectious diseases of humans: dynamics and control, Oxford university press, 1992.
[27] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical review letters 86 (2001) 3200.
[28] M. Barthélemy, A. Barrat, R. Pastor-Satorras, A. Vespignani, Dynamical patterns of epidemic outbreaks in complex heterogeneous networks, Journal of theoretical biology 235 (2005) 275–288.
[29] R. Pastor-Satorras, A. Vespignani, Immunization of complex networks, Physical Review E 65 (2002) 036104.
[30] Z. Dezső, A.-L. Barabási, Halting viruses in scale-free networks, Physical Review E 65 (2002) 055103.
[31] L. A. Meyers, M. Newman, M. Martin, S. Schrag, Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks, Emerging infectious diseases 9 (2003) 204.
[32] P. M. Sloot, S. V. Ivanov, A. V. Boukhanovsky, D. A. van de Vijver, C. A. Boucher, Stochastic simulation of hiv population dynamics through complex network modelling, International Journal of Computer Mathematics 85 (2008) 1175–1187.
[33] M. Small, D. M. Walker, C. K. Tse, Scale-free distribution of avian influenza outbreaks, Physical review letters 99 (2007) 188702.
[34] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A. Vespignani, Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions, PLoS medicine 4 (2007).
[35] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, L. E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60 (2011) 329–412.
[36] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track covid-19 in real time, The Lancet Infectious Diseases (2020).
[37] L. A. Adamic, E. Adar, Friends and neighbors on the web, Social networks 25 (2003) 211–230.
[38] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology 58 (2007) 1019–1031.
[39] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web., Technical Report, Stanford InfoLab, 1999.
[40] G. Jeh, J. Widom, Simrank: a measure of structural-context similarity, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 538–543.
[41] R. Guns, Link prediction, in: Measuring scholarly impact, Springer, 2014, pp. 35–55.
[42] M. E. Newman, Fast algorithm for detecting community structure in networks, Physical Review E 69 (2004) 066133.
[43] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: International symposium on computer and information sciences, Springer, pp. 284–293.

Contributors

T.C. and L.Z. designed the study. T.C. performed the numerical analysis. T.C. and L.Z. wrote the paper.

Materials & Correspondence

The datasets generated during and/or analyzed in the current study are available from the corresponding author on reasonable request.

Declaration of interests

The authors declare no competing interests.