A novel network-based methodology for analysis of COVID-19 data

Marianna Milano[1,2], Mario Cannataro[1,2]

[1]Department of Medical and Surgical Sciences, University of Catanzaro, Catanzaro, 88100, Italy
[2]Data Analytics Research Center, University of Catanzaro, Catanzaro, 88100, Italy

Abstract

The novel COVID-19 pandemic has posed unprecedented challenges to the society and the health sector all over the globe. Here, we present a new network-based methodology to analyze COVID-19 data measures and its application on a real dataset. The goal of the methodology is to analyze set of homogeneous datasets (i.e. COVID-19 data in several regions) using a statistical test to find similar/dissimilar dataset, mapping such similarity information on a graph and then using community detection algorithm to visualize and analyze the initial dataset. The methodology and its implementation as R function are publicly available at https://github.com/mmilano87/analyzeC19D. We evaluated diverse Italian COVID-19 data made publicly available by the Italian Protezione Civile Department at https://github.com/pcm-dpc/COVID-19/. We considered the data provided for each Italian region in two periods February 24-April 26, 2020 (1^{st} wave), and September 28-November 29, 2020 (2^{nd} wave) and then we compared two periods. Similarity matrices of Italian regions for ten COVID-19 data measures are built by using statistical analysis; then they are mapped to undirected networks. Each node represents an Italian region and an edge connects statistically similar regions. Finally, clusters of regions with similar behaviour were found using network-based community detection algorithms. Experiments depict the communities formed by Italian regions over time and the communities change with respect to the ten data measures and time.

Keywords

COVID-19, Network Analysis, Communities Detection

1. Introduction

The coronavirus (COVID-19) epidemic started in China, in November 2019 [1] and it has quickly spread in hundreds of countries in the world. COVID-19 disease is a viral infection produced by Sars-CoV-2, a new coronavirus [2]. The high transmissibility, the high level of infectivity [3] and an initial absence of a COVID-19 vaccine caused about a huge number of deceases. In March 2020, World Health Organization (WHO) declared COVID-19 as a serious epidemic. Thousands of cases of COVID-19 are recorded in Italy, and over 20,000 patients have died. In Italy, the first transmission of COVID-19 disease was reported on February, 2020. The outbreak of COVID-19 initiated in the Italian northern regions [4] and it spread rapidly in the rest of regions. Italy was seriously influenced by the epidemic with more than $3,200,000$ infected and more than $115,000$ deaths that recorded at the end of April, 2021.

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy
m.milano@unicz.it (M. Milano); cannataro@unicz.it (M. Cannataro)
0000-0003-1561-725X (M. Milano); 0000-0003-1502-2387 (M. Cannataro)

In this work, we analyze the spread of the disease over time at a regional level by underlining difference among them. To this aim, in [5], we implemented a new methodology that is able to depict COVID-19 data as networks and to apply graph-based methods in order to evidence regions with similar behaviour along the time. In this work, we analyzed the evolution of the communities by considering a longer interval of time with the goal to highlight original clusters of regions with respect to COVID-19 data. We analyze the data released every day from the Italian Civil Protection. At first, we focus on the data provided on the period February 24-April 26, 2020. Then, we point out the period from September 28 to November 29, 2020. In detail, for each region, we considered ten daily provided data (e.g. total cases, deceased, see the rest of measures below) with the aim of assessing the behaviour of Italian regions respect to COVID-19 and highlighting which regions exhibit similar behaviour. To do this, we defined a new methodology to analyze those data, that comprises the four steps: i) Application of statistical test to identify the regions that present similar/dissimilar behaviour respect to COVID-19; ii) Building of similarity matrices; iii) Mapping of matrix of similarity into network whose each node is an Italian region and each edge depicts similarity connections; iv) Identification of communities by applying community detection algorithms.

The fundamental contribution of the paper consists of a new methodology that is able to depict COVID-19 data as networks and to apply graph-based methods in order to evidence regions with similar behaviour. The analysis enables to remarkably present the development of the communities over time, how the communities growth due to joining of regions or the communities reduce due to leaving of regions. This ensures to evaluate the changes of community coherence in relation to different data. The paper is designed as follows: Section 1 discusses the background on community detection on the network, Section 2 introduces the new methodology for network-based analysis, Section 3 discusses the use of our methodology on Italian COVID-19 data, Section 4 concludes the paper.

2. Methodology

We developed a new methodology for network-based analysis, where the input data are a collection of homogeneous measures (e.g. COVID-19 data in similar regions). The analysis pipeline consists of four steps:

1. **Definition of the similarity matrix.** In the first step the identification of similarity benchmark for the analyzed data is performed. Once similarity measure is applied to data, a similarity matrix is built. Let matrix M for dataset k, each (i, j) element represents a value obtained by performing a similarity measure.
   Given an input dataset $D_1 ... D_n$, where each D_i may contain one or more measured data $di_1^K ... di_n^K$, with 1...k time series. The similarity matrix for data $K = d_{ij}^K$ over a time period T is a matrix M where the value $M(i, j)$ is the similarity among the dataset D_i and the dataset D_i i.e. among $\{di_1^K ... di_n^K\}$ and $\{dj_1^K ... dj_n^K\}$; $(i, j) = 1...N$. In this work, the dataset i means the Region i.
2. **Converting similarity matrix to network**. In the second step the building of a similarity network is performed. According to the similarity matrix, a similarity threshold is fixed. Then, each similarity matrix $M(i, j)$ is mapped to a network N, whose nodes are

the Italian regions and the edges connect them when the similarity value among two regions (i,j), resulting in the matrix, exceeds the similarity threshold. Edges are weighted according to the similarity values. The lengths of the edges result inversely proportional to similarity values.

3. **Network analysis over time.** In the third step the building of the network at different time intervals is performed. Starting from consideration that the COVID-19 data present a temporal evolution, for each one, the corresponding networks at diverse time intervals and for an observation period are built. Furthermore, this step enables the visualization of the networks changes at different time.
4. **Community detection.** In the last step the extraction of the communities on the built networks is performed. The choice of an appropriate community detection algorithm is important to achieve the best results. Thus, an algorithm for community detection is applied to identify communities, i.e. groups of regions sharing similarity, on the networks related to different time intervals and for an observation period. Finally, this step enables the visualization of the community at different time changes.

Figure 1 shows the pipeline of methodology.

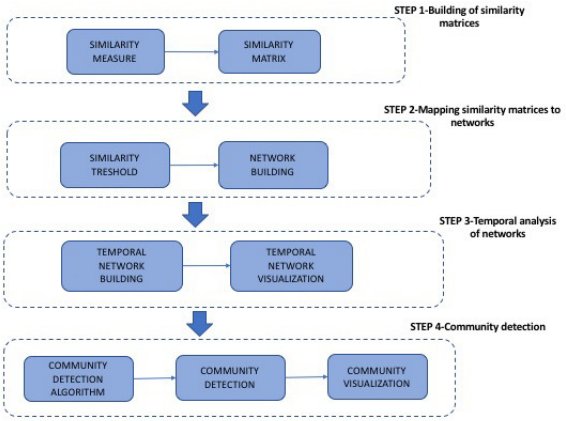


Figure 1: Methodology pipeline

3. Results

Our methodology presents a general design, for which it can be applied to analyze different types of data. In this work, we apply our methodology to analyze Italian COVID-19 dataset. We implement our algorithm using R software [6]. The algorithm is available at https://github.com/mmilano87/analyzeC19D. We chose igraph libraries [7] to perform the network analysis.

3.1. DataSet

We performed the analysis on COVID-19 dataset, released daily by the Italian Civil Protection at https://github.com/pcm-dpc/COVID-19 database. The overall database contains a dataset

for each day, starting from February 24, 2020 and for each Italian region. The regions are: Abruzzo, Basilicata, Calabria, Campania, Emilia, Friuli, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Umbria, Valle d'Aosta, Veneto, plus the autonomous provinces of Bolzano and Trento, for a total of 21 regions. The dataset of Regions in a time point (day) contains the following measurements:

- Hospitalised with Symptoms, relating to the count of COVID-19 patients in the hospital;
- Intensive Care, relating to the count of COVID-19 patients in Intensive Care Units;
- Total Hospitalised, relating to the sum of Hospitalised with Symptoms and Intensive Care measured;
- Home Isolation, relating to the count of subjects in quarantine at home;
- Total Currently Positive, relating to the count of COVID-19 positive subjects;
- New Currently Positive, relating to the daily count of COVID-19 positive subjects;
- Discharged/ Healed, relating to the count of healed or discharged from hospital subjects;
- Deceased, relating to the count of deaths;
- Total Cases, relating to the count of subjects affected by COVID-19;
- Swabs, relating to the count of test swab carried on COVID-19 positive subjects and on suspected COVID-19 positivity subjects.

The data have not been normalized with respect the number of people in a region.

The data occupies 169 Mbytes of memory.

3.2. Building of similarity matrices

We started to build the similarity matrices of each COVID-19 collected data. At first we focus on COVID-19 data related to the February 24-April 26, 2020 period.

Then, we analyze the COVID-19 data related to the September 29-November 24, 2020 period (that, for convenience, we called second observation period).

Initially, we used Pearson's chi-square test to evaluate the distribution of each type of data.

Since *p-value* resulted less than 0.05, we chose to applied non-parametric test for for subsequent analysis.

So, we recurred to the use of the Wilcoxon Sum Rank test to analyze the same type of data on the observation period, consisting of nine weeks and then, for single time intervals. The Wilcoxon test enables to assess the difference among to conditions when the samples are correlated. Thus, we used the Wilcoxon test to compare the Italian regions with the aim to evidence statistically similar distributions among them. After that, we constructed ten similarity matrices for all COVID-19 data and for each time interval, that report the statistical comparison among a couple of regions.

In detail, the (i,j) value of the matrix, related to data k (for example deceased data), describes the *p-value* of the Wilcoxon test achieved by applying the test on a given measure (e.g. the number of deceased) of the region i versus the region j in a given time interval. Lower *p-value* implies that two regions are different according to that measure. Otherwise, higher *p-value* implies that regions show a similarity according to that measure. We considered the conventional significance threshold of 0.05, and for this reason we built similarity matrices that contain only $p - values >= 0.05$. We mapped the $p - values < 0.05$ equal to zero. So, we considered the *p-value* as a measure of similarity.

3.3. Mapping similarity matrices to networks: Network building

We mapped each similarity matrix M(i,j) into a network N [8]. The nodes are the Italian regions, and the edges link two regions (i,j) when the *p-value* is greater than the threshold, otherwise (p-value < 0.05) none edge is added. Each edges is weighted according to *p-value* resulting from Wilcoxon test. In this way, the edge length results inversely proportional to similarity.

3.4. Network Visualization over time

We performed a temporal analysis by building ten networks related to the data measures (Hospitalised with Symptoms, Intensive Care data, Total Hospitalised, Home Isolation, Total Currently Positive, New Currently Positive, Discharged/ Healed, Deceased, Total Cases, Swab) by considering the period February 24 to April 26 (the observation period) and then, by building the networks of the same data at different time intervals. After that, we built the networks related to the second observation period and single weeks.

3.5. Temporal evolution of Communities

For each network related to different time intervals, our goal was the identification of regions that construct a community according to their similarity. For this aim, we used Walktrap community finding algorithm [9]. Walktrap is able to identify subgraphs with high density, i.e. communities, in a network through random walks. We selected Walktrap community detection algorithm because it outperforms other methods as discussed in [10]. For example, Figure 2 represents the evolution of Deceased Network Communities in the first observation period and Figure 3 the evolution the second observation period.

The results show that diverse data show dissimilar communities. In fact, each network related to different Italian COVID-19 data presents different detected communities. Furthermore, the results highlight that communities formed by regions evolves according to the diverse data. In particular, the different community structure grow due to joining of regions or the communities reduces due to leaving of regions. This enables to evaluate the changes of community according to to different data [5].

Figure 2 (a) presents the mined communities of Deceased Network at the end of the first week. The detected communities consist of: a considerable community formed by Emilia, Piemonte, Liguria, Campania, Abruzzo, Puglia, Valle d'Aosta, Umbria, Calabria, Sicilia, Campania, Trento, Lazio, Sardegna, Bolzano, Basilicata and Molise, Friuli; a single community formed by Lombardia: a single community represented by Veneto; and a single community consisting of Marche and Toscana.

By considering the networks built in the second observation period it is possible to note that the topology evolves from a sparse to a dense structure and this reflects on the discovered communities. For example, in Deceased Network, each region forms single community in single week and after three weeks, whereas, after five weeks, three communities are composed by two regions (Figure 3 (c)): (i) Marche and Lazio, (ii) Campania and Abruzzo, (iii) Sicilia and Friuli. After seven weeks (Figure 3 (d)), the Italian regions form single communities with the exception of three communities formed respectively by Umbria and Calabria, Campania and Abruzzo, Lazio and Marche and a community formed by three regions, Friuli, Sicilia, Trento.

4. Conclusion and Future Work

In conclusion, the aim of this work is to provide a network-based representation of the COVID-19 measures in order to evaluate the region behaviour on the basis of different data provided by Civil Protection. For this, we detected similar region related to COVID-19 data, and we mapped them in different networks. Then, we conducted a network-based analysis to identified communities of regions with similar behaviour. The present methodology is general and can be applied to the connection of data varying over time. Currently, the new methodology is available as R function. As future work, we plan to implement the methodology as an R package.

References

[1] Z. Wu, J. M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention, Jama (2020).

[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in china, 2019, New England Journal of Medicine (2020).

[3] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The Lancet 395 (2020) 497–506.

[4] A. Lai, A. Bergna, C. Acciarri, M. Galli, G. Zehender, Early phylogenetic estimate of the effective reproduction number of sars-cov-2, Journal of medical virology (2020).

[5] M. Milano, M. Cannataro, Statistical and network-based analysis of italian covid-19 data: Communities detection and temporal evolution, International journal of environmental research and public health 17 (2020) 4182.

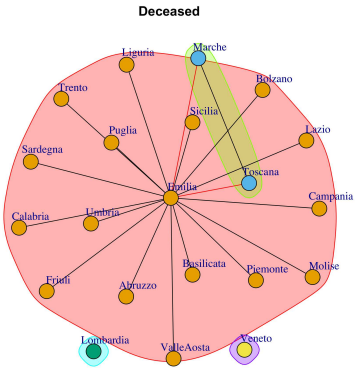
[6] M. Milano, Computing languages for bioinformatics: R, Gribskov, M., Nakai, K. and Schonbach, C. (eds.), Encyclopedia of Bioinformatics and Computational Biology,Elsevier, Oxford: Elsevier. 1 (2019) 889–895.

[7] G. Csardi, T. Nepusz, et al., The igraph software package for complex network research, InterJournal, complex systems 1695 (2006) 1–9.

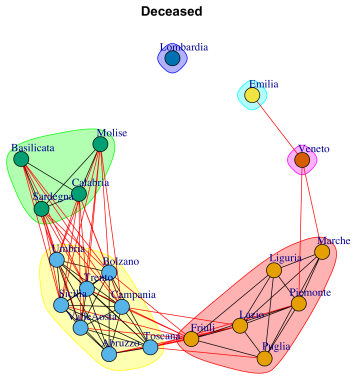
[8] G. Agapito, P. H. Guzzi, M. Cannataro, Challenges and opportunities for visualization and analysis of graph-modeled medical data, 2017. URL: https://doi.org/10.20944/preprints201710.0018.v1. doi:10.20944/preprints201710.0018.v1.

[9] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: International symposium on computer and information sciences, Springer, 2005, pp. 284–293.

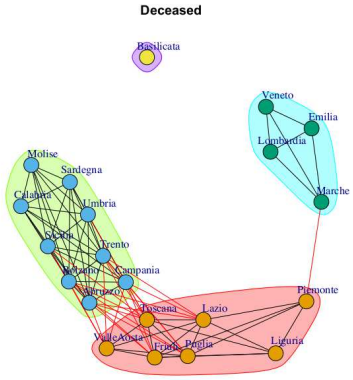
[10] F. B. de Sousa, L. Zhao, Evaluating and comparing the igraph community detection algorithms, in: 2014 Brazilian Conference on Intelligent Systems, IEEE, 2014, pp. 408–413.



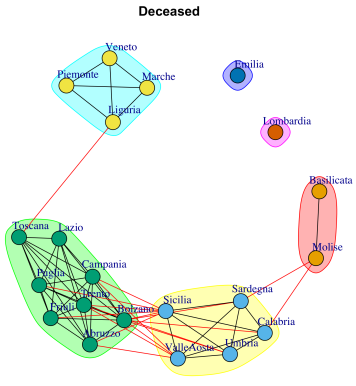
(a) First Week



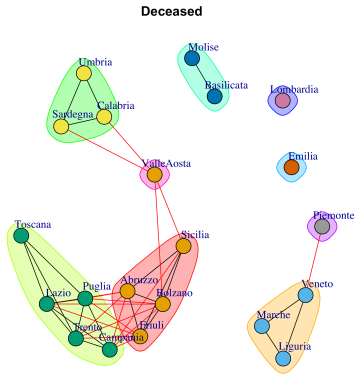
(b) After Three Weeks



(c) After Five Weeks

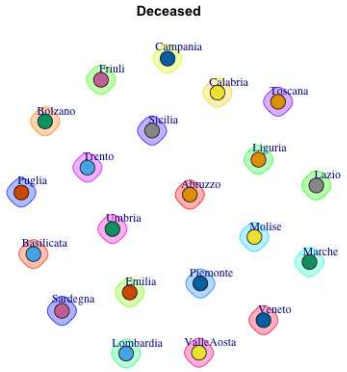


(d)After Seven Weeks

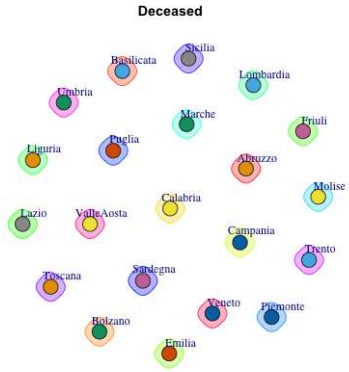


(e)After Nine Weeks

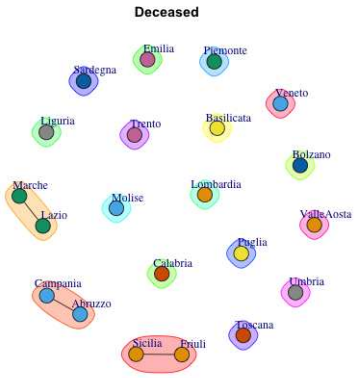
Fig. 2. Evolution of Deceased Network Communities in the first observation period.



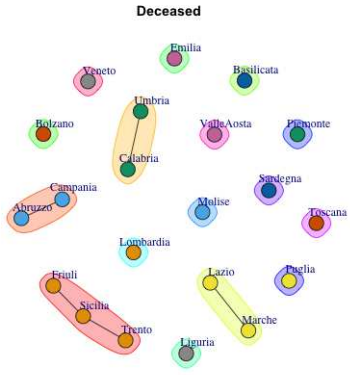
(a) First Week



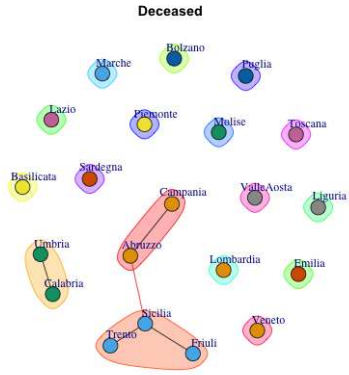
(b) After Three Weeks



(c) After Five Weeks



(d)After Seven Weeks



(e)After Nine Weeks

Fig. 3. Evolution of Deceased Network Communities in the second observation period.