A New Parallel Methodology for the Network Analysis of COVID-19 Data

Giuseppe Agapito\textsuperscript{1,3}, Marianna Milano\textsuperscript{1,2(✉)}, and Mario Cannataro\textsuperscript{1,2}

\textsuperscript{1} Data Analytics Research Center, Magna Græcia University, 88100 Catanzaro, Italy
\{agapito,m.milano,cannataro\}@unicz.it
\textsuperscript{2} Department of Medical and Surgical Sciences, Magna Græcia University, 88100 Catanzaro, Italy
\textsuperscript{3} Department of Legal, Economic and Social Sciences, Magna Græcia University, 88100 Catanzaro, Italy

Abstract. Coronavirus disease (COVID-19) outbreak started at Wuhan, China, and it has rapidly spread across the world. In this article, we present a new methodology for network-based analysis of Italian COVID-19 data. The methodology includes the following steps: (i) a parallel methodology to build similarity matrices that represent similar or dissimilar regions with respect to data; (ii) the mapping of similarity matrices into networks where nodes represent Italian regions, and edges represent similarity relationships; (iii) the discovering communities of regions that show similar behaviour. The methodology is general and can be applied to world-wide data about COVID-19. Experiments was performed on real datasets about Italian regions, and they although the limited size of the Italian COVID-19 dataset, a quite linear speed-up was obtained up to six cores.

Keywords: COVID-19 · Network analysis · Parallel computing

1 Introduction

The global pandemic is caused by a new coronavirus named \textit{Severe Acute Respiratory Syndrome CoronaVirus 2} (SARS-CoV-2) [16], which was first discovered in December 2019 in China [14]. In six months, COVID-19 has spread to more than two hundred countries, infected millions of people and it caused about tens of millions of deaths. COVID-19 has become a global pandemic not only because SARS-COV-2 is new without an effective treatment, but also because it is transmissible from person to person. On March 12, 2020, the World Health Organization (WHO) announces COVID-19 outbreak as a pandemic.

COVID-19 has been recognized in Italy starting from January 31, 2020, [7]. The spread of the disease started from the northern regions of Italy, Lombardy and Veneto on February 21, 2020. From the northern regions of Italy, the disease spread very quickly to the nearest regions and then to the rest ones.
The aim of this study consists of providing a network-based representation of the behaviour of Italian regions with respect to COVID-19 outbreak. To do this, we design an analysis pipeline to model Italian COVID-19 data, daily provided by the Italian Civil Protection, as networks and to perform network-based analysis. We collected data in the period from February 24th to June 7th, 2020. At first, for each type of data, we evaluate the similarity among pair of regions by using a statistical test, i.e. Wilcoxon Sum Rank Test, and according to this, we built similarity matrices (one for each Italian COVID-19 data measure released by Italian Civil Protection). To improve the computation of the similarity matrices, we implemented a parallel methodology. The parallelization of the similarity-matrices calculation is a problem of allocating independent tasks to parallel processors.

In literature, different works recur to network-modelling to analyze COVID-19 data, and most of them recur to network-based representation of data for the application of predictive models. For example, Reich et al. [11] implemented the COVID-19 spread by using SEIRS (Susceptible-Exposed-Infectious-Recovered-Susceptible) agent-based model on a network; Kuzdeuov et al. [6] implemented a network-based stochastic epidemic simulator that models the movement of a disease through the SEIR states of a population; Kumar [5] presents a network-based model for predicting the spread of COVID-19.

To the best of our knowledge, our work is the first study that provides a network-based representation and visualization of COVID-19 data at the regional level and applies network-based analysis to discover communities of regions that show similar behaviour by using a parallel methodology.

The first step of the methodology involves the calculation of the similarity matrix between each pair of regions. This is an easy target for coarse-grained parallelization since all elements of the similarity matrix are independent. We obtain the tasks by considering the input data as a matrix $A_{n \times m}$ with $n$ rows and $m$ columns, from which to extract the squared-blocks $SB$ (i.e., squared sub-matrices between pair of regions) to compute the similarity matrices.

The second step consists in converting the SBs in a network where the nodes represent the Italian regions, and the edges connect statistically similar regions. Finally, we extracted subgroups of regions that form communities based on similarity point of view. The proposed methodology is targeted for multiple CPUs/Cores shared-memory machines.

The main contributions of the paper are: i) a parallel preprocessing methodology to improve the multiple pair-wise comparison between Italian regions, ii) a network-based representation of COVID-19 diffusion similarity among regions and, iii) a graph-based visualization to underline similar diffusion regions, that have a similar diffusion pattern of the disease.

The rest of the paper is organized as follows: Sect. 2 discusses the background on community detection, Sect. 3 presents the implemented pipeline to analyze Italian COVID-19 data, Sect. 4 discusses the results. Finally, Sect. 5 concludes the paper.
2 Background

As a modelling framework, the complex network model has been applied in different fields such as biology, computer science, communication. Once modelled, the network is analysed using some of the many algorithms designed for graph mining. In general, the networks are featured by a heterogeneous structure with specific properties. In particular, the structure presents a heterogeneous distribution of edges that identified the presence of the community. A community presents a group of nodes high densely interconnected respect to the rest of the network [4]. Regardless of the nature of the network, community conveys very important information for the understanding of structural properties. So, community detection in networks is one of the most popular topics in network analysis. In literature, there are many different community detection algorithms.

For example, WalkTrap [10] is a hierarchical clustering algorithm that applies a distance measure based on random walks. Initially, WalkTrap computes the distances between all adjacent nodes in the network. Then, it starts with a node and randomly selects a neighbour of the current node; it merges them in a community, and it updates distances between communities. The idea is that short random walks tend to stay in the same community.

MarkovCluster algorithm [13] works by simulating a stochastic (Markov) flow in a weighted graph, where each node is a data point, while the adjacency matrix stores the edge weights.

Fast Greedy algorithm [2] uses a basic greedy approach [15] starting by single nodes and it joins pairs of ones to form communities.

Louvain [1] includes a community aggregation step to improve communities detection process. The algorithm joins a node with each one of its neighbours community according to the increasing of modularity, otherwise the node stays in its original community.

Spinglass [12] algorithm is based on physical spin glass models. The algorithm aims to find ground state of a spin glass model on the basis that the edges should connect nodes of the same spin state.

3 Parallel Analysis Pipeline

We designed the analysis pipeline with the goal to investigate clusters of Italian Regions with similar behaviour with respect to data provided by the Italian Civil Protection.

The analysis pipeline includes the following steps:

1. Building of similarity matrix. In the first step, the similarity matrix is built. The matrix enters the similarity among pairs of regions respect to an Italian COVID-19 data measure. The similarity is computed by performing the Wilcoxon Sum Rank statical test. Thus, the \((h, k)\) value of the similarity matrix \(M\) for data \(A\), e.g., Intensive Care data, is the \(p\)-value of statistical test obtained by applying the test on the Intensive Care measures of region \(h\) with respect to region \(k\). Lower \(p\)-value implies that regions are dissimilar.
with respect to that measure. Otherwise, higher \( p\)-value implies that regions are similar with respect to that measure. We used the significance threshold of 0.05, thus matrices report only the \( p\)-values such that: \( p\)-values \( \geq \) 0.05, while \( p\)-values \( < \) 0.05 are mapped to zero.

2. Mapping similarity matrices to networks. The second step consists of the building networks starting from the similarity matrices. We map each matrix \( M(h, k) \) to a network \( N \), where nodes represent the Italian regions and an edge connects two regions \( (h, k) \) if the \( p\)-value in the similarity matrix is greater than the significance threshold of 0.05. Edges are weighted with the \( p\)-value.

3. Community detection. The third step consists of the detection of communities on the network by applying an appropriate community detection algorithm. For each network, we extracted subgroups of regions that form a community on the basis of similarity point of view.

Figure 1 shows the main steps of the parallel analysis pipeline.

4 Results

We applied the designed pipeline to analyze the Italian COVID-19 data by considering the period from February 24 to 7 June.

4.1 Input Dataset

The present analysis was carried on the Italian dataset on COVID-19 available at the https://github.com/pcm-dpc/COVID-19 database, provided by the Italian Civil Protection. The dataset consist of the following data collected daily:

- Hospitalised with Symptoms, the numbers of hospitalised patients that present COVID-19 symptoms;
- Intensive Care, the numbers of hospitalised patients in Intensive Care Units;
- Total Hospitalised, the total numbers of hospitalised patients;
- Home Isolation, the numbers of subjects that are infected and in isolation at home;
- Total Currently Positive, the numbers of subjects that are coronavirus positive;
- New Currently Positive, the numbers of subjects that are daily coronavirus positive;
- Discharged/Healed the numbers of subjects that are healed from the disease;
- Deceased, the numbers of dead patients;
- Total Cases, the numbers of subjects affected by COVID-19;
- Swabs, the numbers of swab test carried on positive subjects and on subjects with suspected positivity.

The data are daily provided for each Italian region.

4.2 Parallel Building of Similarity Matrices

In order to build similarity matrices for Italian COVID-19 data, we performed the Wilcoxon Sum Rank Test. The analysis is performed by using R software [8]. We performed the Wilcoxon test to compute a pair-wise comparison among regions with the goal of evidence statistically similar distributions among them. The pair-wise similarity computation can be executed in parallel since it is an embarrassingly parallelizable task. The problem can be defined as follows: given a matrix $A_{(n \times m)}$ with $n$ rows and $m$ columns, where each element is denoted as $a_{ij}$ with $1 \leq i \leq n$ and $1 \leq j \leq m$. The matrix $A$ is virtually split in $p$ Square-Blocks $SB$, where $p$ is the number of available cores, and it is used to balance the workload among the processors/cores available. $A$ $SB$ is built for each couple of regions, and for all the available COVID-19 data (Hospitalised with Symptoms, Intensive Care data, Total Hospitalised, Home Isolation, Total Currently Positive, New Currently Positive, Discharged/Healed, Deceased, Total Cases, Swab). In particular, the main steps necessary to build the similarity matrices are:

1. Partitioning the matrix $A$ in $p$ Square Blocks SB, to balance the workload among the cores/CPUs available.
2. create independent threads whom is assigned a SB, so that each slave can independently compute its part of the similarity matrix;
3. each computed SB is added to the similarity matrix.

The parallel methodology is currently implemented as a multi-threaded Python application using the threading library. The experiments were performed on a workstation equipped with a Pentium i7 2.3 GHz CPU, 16 GB RAM and a 512 GB SSD disk.

The speed-up ($S$) is defined as the ratio of the time taken using a single processor ($T(1)$) over the time measured using $n$ processors ($T(n)$) (see Eq. 1).

$$S(n) = \frac{T(1)}{T(n)}$$  \hspace{1cm} (1)

Figure 2 reports the speed-up obtained by analyzing italian COVID-19 data using 1, 2, 4, and 6 cores, respectively.

Analyzing Fig. 2, it is worth noting that the speed up tends to decrease by increasing the number of computational cores. This is due to the low volume of
available COVID-19 data to analyze; in fact, increasing the number of computational cores, the time spent to communicate and synchronize between computational cores, is more expensive than the time necessary to perform computation, leading to low values of speed-up.

4.3 Parallel Mapping Similarity Matrices to Networks

The nodes of the networks are the Italian regions, and the edges link two regions (nodes) with similar trend according to significance level (p-value > 0.05) obtained from Wilcoxon test, otherwise (p-value < 0.05) there is no connection among nodes. The network analysis is performed in parallel by using the igraph library [3]. Results show that according to the type of data, a significant difference exists (p-value less than 0.05) among some regions while for others, it is possible to highlight statistically similar distributions.

4.4 Community Detection

With the goal to identify which regions form a community from the similarity point of view, we applied Walktrap community finding algorithm [10] on the networks, to identify densely connected subgraphs. The extracted communities from Italian COVID-19 networks in the observation period (from February 24th to June 7th, 2020) are reported in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.
Fig. 3. Communities in Hospitalised with Symptoms Network.

Fig. 4. Communities in Intensive Care Network.

Fig. 5. Communities in Total Hospitalised Network.
Fig. 6. Communities in Home Isolation Network.

Fig. 7. Communities in Total Currently Positive Network.

Fig. 8. Communities in New Currently Positive Network.
Fig. 9. Communities in Discharged/Healed Network.

Fig. 10. Communities in Total Cases Network.

Fig. 11. Communities in Deceased Network.
The results highlight different community structures consisting of groups of regions with similar trends with respect to different data. In fact, each network related to different Italian COVID-19 data presents different detected communities. Furthermore, the regions that form the communities vary according to the diverse data. For example, Lombardia forms a single community in Hospitalised with Symptoms Network (Fig. 3), Total Hospitalised Network (Fig. 5), Total Currently Positive Network (Fig. 7), New Currently Positive Network (Fig. 8), Discharged/Healed Network (Fig. 9). Otherwise, in Intensive Care Network (Fig. 4) and Swabs Network (Fig. 12), Lombardia forms a community together with Veneto. In Home Isolation Network (Fig. 6), a community is composed by Lombardia, Emilia, Marche. In Total Cases Network (Fig. 10) and Deceased Network (Fig. 11), Lombardia forms a community with Lombardia, Emilia, Marche and Veneto. This means that the Italian regions behaved differently with respect to ten different data provided by civil protection. Further analysis about communities in Italian COVID-19 data and their temporal evolution can be found in [9], where a sequential version of this methodology has been used.

5 Conclusion

We proposed a new parallel methodology for network-based representation of COVID-19 similarity among regions and their graph-based visualization, with the aim to underline similar diffusion regions. We identified similar Italian regions with respect to the available COVID-19 data, and we mapped these ones in different networks. Finally, we performed a network-based analysis to discover communities of regions that show similar behaviour. The experiments performed on real datasets show good speed-up. As future work we plan to implement an R package that incapsulates the presented pipeline.
References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
2. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
3. Csardi, G., Nepusz, T., et al.: The igraph software package for complex network research. InterJournal Complex Syst. **1695**(5), 1–9 (2006)
4. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
5. Kumar, A.: Modeling geographical spread of COVID-19 in India using network-based approach. medRxiv (2020). https://doi.org/10.1101/2020.04.23.20076489
6. Kuzdeuov, A., et al.: A network-based stochastic epidemic simulator: controlling COVID-19 with region-specific policies. medRxiv (2020). https://doi.org/10.1101/2020.05.02.20089136
7. Lai, A., Bergna, A., Acciarri, C., Galli, M., Zehender, G.: Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. J. Med. Virol. (2020). https://doi.org/10.1002/jmv.25723
8. Milano, M.: Computing languages for bioinformatics: R. In: Gribskov, M., Nakai, K., Schonbach, C. (eds.) Encyclopedia of Bioinformatics and Computational Biology, vol. 1, pp. 889–895. Elsevier, Oxford (2019)
9. Milano, M., Cannataro, M.: Statistical and network-based analysis of Italian COVID-19 data: communities detection and temporal evolution. Int. J. Environ. Res. Public Health **17**(12), 4182 (2020)
10. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Yolum, I., Güngör, T., Gürgen, F., Özturan, C. (eds.) ISCIS 2005. LNCS, vol. 3733, pp. 284–293. Springer, Heidelberg (2005). https://doi.org/10.1007/11569596_31
11. Reich, O., Shalev, G., Kalvari, T.: Modeling COVID-19 on a network: superspreaders, testing and containment. medRxiv (2020). https://doi.org/10.1101/2020.04.30.20081828
12. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**(1), 016110 (2006)
13. Van Dongen, S.: Graph clustering via a discrete uncoupling process. SIAM J. Matrix Anal. Appl. **30**(1), 121–141 (2008)
14. Wu, Z., McGoogan, J.M.: 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). https://doi.org/10.1001/jama.2020.2648
15. Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. **6**, 30750 (2016)
16. Zhu, N., et al.: A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. (2020). https://doi.org/10.1056/NEJMoA2001017