A Graph-based Methodology for Tracking Covid-19 in Time Series Datasets

2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) | 978-1-7281-6921-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICECOCS50124.2020.9314516

Zakariyaa Ait El Mouden
Software Engineering & Information Systems Engineering, Faculty of Sciences and Techniques, Errachidia, Moulay Ismail University, Meknes, Morocco.
mouden.zakariyaa@outlook.com

Abdeslam Jakimi
Software Engineering & Information Systems Engineering, Faculty of Sciences and Techniques, Errachidia, Moulay Ismail University, Meknes, Morocco.
ajakimi@yahoo.fr

Rachida Moulay Taj
Operational Research & Computer Science, Faculty of Sciences and Techniques, Errachidia, Moulay Ismail University, Meknes, Morocco.
rachidamoulaytaj@gmail.com

Moha Hajar
Operational Research & Computer Science, Faculty of Sciences and Techniques, Errachidia, Moulay Ismail University, Meknes, Morocco.
moha_hajjar@yahoo.fr

***Abstract*—Since its first appearance in December 2019, Covid-19 has become a wide field of scientific research. Starting from biology to bioinformatic solutions, Artificial Intelligence has contributed in turn as a powerful tool for tracking and predicting the outbreak of Covid-19 using different types of datasets. Chest X-ray images are widely used in computer vision applications and Time series datasets are used for predicting the spread of the novel coronavirus. Graph analytics is a recent field of study that links the mathematical definition and operations of a graph to its application in computer science as a complex data structure, this combination has played a critical role in making graph-based applications present in different fields. One of the most powerful graph analytics is community detection which is an intelligent and unsupervised grouping of a set of graph structured data using the similarity between them. The aim of this work is to highlight the importance of graph-based algorithms in tracking Covid-19 using time series datasets, our work will also focus on Spectral Clustering (SC) as a community detection approach to extract clusters from the input datasets. Further applications are needed in order to validate the proposed theoretical approach.**

Keywords— Covid-19, Tracking, Graph analytics, Communities detection, Spectral clustering.

I. Introduction

Since its first appearance in late December 2019, Covid-19 has become a worldwide concern, this coronavirus disease is caused by a virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1].

Covid-19 was initially discovered in Wuhan China, after three months of its appearance, in March, 2020, the novel coronavirus was identified as a pandemic because of its fast outbreak around the world. To date (19th September 2020), the number of Covid-19 cases has reached 30,941,513 including 960,275 deaths and 22,549,569 recovers. The number of active cases is 7,431687 including 61,459 characterized as critical cases [2]. According to Worldometers, the United States are still in the top of the table with almost 6 million cases since the first appearance of the virus on January, 19 2020 [3].

In Morocco, the first infected case with covid-19 was declared in March, 2 2020 in Casablanca city [4]. While writing this paper, the number of covid-19 cases in Morocco has reached 99,816 including 1,795 deaths. The number of active cases in Morocco is up to 19,013 including 274 characterized as critical/serious cases.

The majority of Covid-19 cases has similar symptoms to normal influenza such as fever, sneezing, cough and other respiratory disorders. The incubation period from the infection to the appearance of the symptoms can take from 2 days to 14 days in the most cases. The normal symptoms of Covid-19 in addition to its long incubation period play a critical role in the spread of this novel coronavirus and make it very difficult to control in the most infected countries.

Classification techniques such as multiclass classifiers are widely used to classify images into a predefined set of classes, also called output layers in Neural networks [5], those techniques give good results for Covid-19 detection such as the application to X-ray images [6, 7] or the classification of the infected areas by regions especially using Length Short-Term Memory (LSTM) networks [4, 8]. While, clustering techniques are widely used for grouping data into unknown set of clusters according to their similar behaviors such as the behavior of the Covid-19's outbreak in function of time and other factors [9, 10].

This contribution is an unsupervised learning methodology of an intelligent grouping of countries according to the similar spreading of Covid-19, taking in consideration a set of factors such as the population and other factors that affect the behavior of the outbreak. The application is developed using spectral clustering which is an unsupervised learning algorithm for data modelled by graphs [11-13].

The rest of the paper is outlined as follows; Section II gives a background of some related works, Section III introduces briefly the community analytics in graphs and presents some of its applications, Section IV details the different steps of the proposed methodology. Finally, Section V gives a conclusion and highlights some future contributions.

II. Background

Different machine learning models were applied to analyze the propagation of the novel coronavirus; Truncated Inception net [14] is a deep learning convolutional neural network model for screening chest X-ray images datasets and classify those images to Covid-19 and non-Covid-19, the obtained results were used to track the outbreak of coronavirus. In the same area, the authors of [15] proposed a convolutional neural network CapsNet to detect Covid-19 disease in chest X-ray datasets, the author used the dataset of Cohen [16] which contains images of Covid-19 patients from different countries.

Another contribution is [17], where the authors propose a fractional-order model to analyze the outbreak of Covid-19 using collected data of confirmed cases in Italy and published by the World Health Organization (WHO).

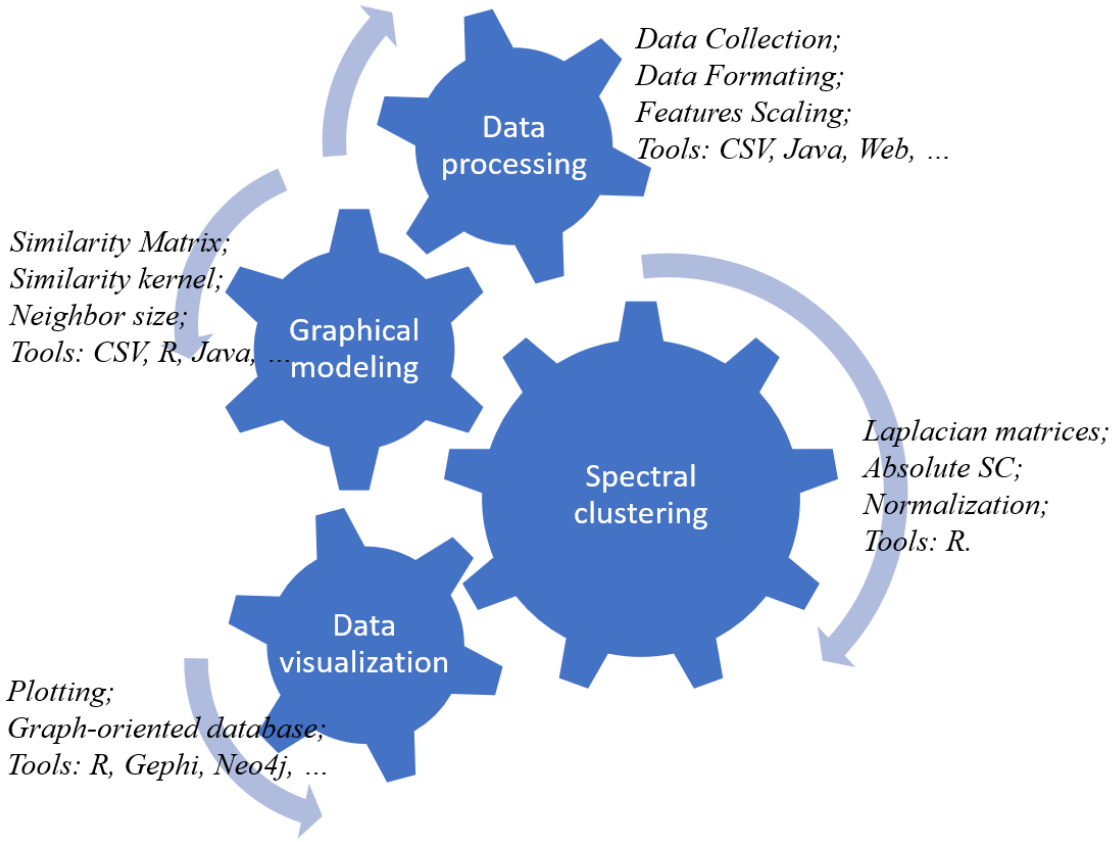

Fig. 1 Process overview.

In [18], the authors presented their data-driven neural network to analyze the effect of quarantine and isolation in controlling the spread of Covid-19, the experiments focuses on four locations: Wuhan, Italy, South Korea and the United States of America. For each country, the authors provided two models, the first is without quarantine measurements and the second is when the quarantine and the isolation are applied strictly, the model showed good results especially for Wuhan city where data were available earlier in comparison to other regions.

Time series are applied to predict Influenza-like Illness and give highlights on the current epidemiological situation in each country. The authors of [19] proposed an LSTM based neural network to forecast time series data and compare the results of their model to state-of-the-art approaches.

In [20], the authors give a review on the most used tools for tracking Covid-19 in regional distribution, the contribution highlights the role of the Geographical Information Systems and maps in decision making, planning and community mobilization.

III. Community Analytics

Community analytics is the main subfield of graph analytics. Starting from a heterogeneous individual, communities' detection is an output set of clusters where each cluster is supposed to group individuals sharing similar properties. Clustering is not unique, as it may change according to the used similarity kernel or the normalization function in preprocessing or even other factors, but we can have a better clustering when we compare between different outputs as there is some metrices to evaluate the quality of the generates clusters or communities.

Let's define:

G: A connected graph;

n: The order of the graph (the number of its nodes);

c: A cluster;

n_c: The number of nodes in a cluster c;

$e_{int}(c)$: The number of edges in a cluster c (internal edges);

$e_{ext}(c)$: The number of edges that links nodes from the cluster c with nodes from other clusters (inter-cluster edges);

The intra-cluster density σ_{int} is defined as:

$$\sigma_{int} = \frac{e_{int}(c)}{n_c(n_c - 1)/2}$$

Also, the inter-cluster density σ_{ext} is defined as:

$$\sigma_{ext} = \frac{e_{ext}(c)}{n_c(n - n_c)}$$

Community detection of clustering can be summarized as a multi-objective function that maximizes the intra-cluster density (σ_{int}) and minimizes the inter-cluster density (σ_{ext}).

Spectral clustering is one of the main graph clustering techniques, it consists of computing the Laplacian matrix of the input graph using its Adjacency matrix and its Degrees matrix. Then, the eigenvalues and the eigen vector are extracted from the calculated Laplacian matrix. Finally, a k-means clustering is applied to extract the output clusters. Depending on which Laplacian matrix is used, we distinguish between three types of SC; *i) Unnormalized SC*, which is

based on the minimization of the number of inter cluster edges and the maximization of the number of inter-cluster edges, this condition is not sufficient in most cases, especially with heterogeneous data [21]. *ii) Normalized SC*, also called normalized cuts, it is based on the use of the normalized version of Laplacian matrix and widely used for image segmentation [22]. *iii) Absolute Normalized SC* [23] is a version of SC that overcomes the limits of the previous algorithms, and it is used for high dimensional data which makes it a good solution to face the challenges of Big data, especially the volume and the variety.

The Louvain algorithm [24] is used for dynamic communities' detection, it starts by creating clusters with single nodes. Then for each vertex, the algorithm puts joins the vertex to its neighbor and calculate a gain function in order to detect each neighbor is similar to the vertex. The algorithm repeats the process until building stable communities, in general it takes a low number of iterations to converge.

A recent graph-based approach is Graph Neural Networks (GNN) [25, 26] which is a combination between the existing definition on Neural Networks and the graph data structure. This model can process different types of complex graphs such as the recursive graphs, and gives good results for community detection in comparison with the state-of-the-art algorithms.

IV. Methodology

The proposed methodology consists of four main processes (See Fig. 1); starting from data processing where we describe the way in which data were collected, organized and analyzed. Second, the graphical modelling which the transformation of the collected data as a graph-oriented model. Third, the spectral clustering process is applied to the built graph in order the extract the existing communities. Finally, the visualization process regroups all the tools and ways to describe data visually.

A. Data processing

Data processing is the shared process by almost all the machine learning model. Recently, researches started to focus more on the input data as a key to end with good results, as even having a good model, badly distributed data can lead to poor results.

Data collection, this study collected the time series data of the Covid-19 cases, deaths, recovers for each country from the first day of the appearance of the virus to August, 31 2020 (Table 1). Johns Hopkins University provides time series data available in github [27].

TABLE I. Dataset Features

Feature	*Type*	*Description*
Country	Text	The unique code of the country
Continent		The code of the country
Population	Number	IR*+
IFR	Number	The Infection Fatality Rate, which is the number of deaths divided by the number of the confirmed cases
Tests'	Vector of integers	Number of medical tests for each day
Confirmed'		Number of Covid-19's confirmed cases
Deaths'		Number of deaths caused by Covid-19
Recovers'		Number of recovers

Data formatting, data were formatted in order to fit as input for our model, the existing dataset has a high redundancy amount of information and it is important to reduce the redundancy in order to help the model to run faster. Our solution was an object-oriented representation where each country is mentioned as a single example with different vectors to represent the time series data.

Features scaling, our proposition is a mean normalization in order to have similar range of values close to the range [-1, 1], the normalization process helps the features to have similar weights and then similar impact on the model convergence, but positive coefficient can be added to make the impact of some features stronger than other features, such as the impact of the number of cases in comparison to the impact of the country feature. A mean normalization is defined as:

$$v'_i = \frac{v_i - mean(V)}{max(V) - min(V)}$$

Where:

v_i is the i^{th} value of the feature v;

v_i' is the normalized value of v_i;

V the set of the values of the feature v;

$max(V)$ and $min(V)$ are the maximal value and the minimal value of the feature respectively.

B. Similarity graph

Similarity graphs are built from the final version of data in data processing. As countries are represented by objects, we calculate the similarity between each pair or country objects. The Gaussian similarity kernel was chosen:

$$s_{ij} = \mathrm{e}^{-\frac{d_{ij}^2}{2\sigma^2}}$$

Where:

d_{ij} is the Euclidean distance between the pair of countries i and j;

σ^2 is the size of the neighbor used as a scaling parameter.

The similarity graph is defined as $G = \{V, E\}$;

Where:

V is the set of vertices; each data point is represented by a vertex;

E is the set of edges between the elements of V, the weighs of the edges are the similarity values.

C. Spectral clustering

The choice of SC is justified by its compatibility with data modeled by graphs and also its ease of implementation. For experiments we used the Absolute SC with an unsupervised chose of the number of clusters, more details about this algorithm can be found in our previous contribution [28].

Algorithm. Absolute SC with unsupervised choice of k number of clusters

Input: Similarity matrix S and n the order of the graph.

1. Extract the adjacency matrix W and the degrees matrix D, both of size nxn;
2. Normalize the values of W using mean normalization and built the matrix A;
3. Compute the Absolute Laplacian matrix using A;
$$L_{abs} = D^{-1/2} A D^{-1/2}$$
4. Let w be the set of the eigenvalues of L_{abs};
5. Compute in w_{abs} the absolute value of each element of w (w_{abs} = abs(w));
6. Extract the k eigenvectors $u_1,\ldots, u_k$ associated to the k largest eigenvalues of w_{abs};
7. Let U be a matrix of size nxk containing the vectors $(u_i)_{i=1,\ldots,k}$ as columns;
8. Cluster the y_i a line vectors in U using k-means algorithm into k clusters $C_1, \ldots, C_k$.

Output: Clusters $C_1,\ldots, C_k$

D. Vizualization

In this part, visualization of the different communities is discussed. The choice of the used tools in this study is justified by the nature of the dataset. From CSV files to graphs many solutions are applied to visualize the output of the model.

R Studio provides a set of packages to manipulate data modelled by graphs and matrices;

igraph is a package that allows to build a graph from difference source of data such as CSV files, Pajek files or complex networks. It also provides functions to manipulate different types of graphs, this study focuses on the implementation of undirected and weighted graphs. gplots is a visualization package that provides displays with different layouts and parameters to customize the output figures.

matrixcalc regroups the different functions to deal with matrix calculations, it also contains functions to build Laplacian matrices which facilitate the implementation of spectral clustering in R.

Neo4j is a graph database management system classified as a NoSQL system, this datastore is provided to create and manipulate graph-based systems and it is also a good visualization solution as it support reading CSV files.

Gephi is an additional solution for creating graph models from CSV or XML files, its additional advantage is its ability to visualize dynamic systems which is not possible using previous solutions.

V. Conclusions

In this paper, we introduced a new graph-based methodology for tracking Covid-19 in time series datasets, our four steps methodology consists of data processing, graphical modeling, spectral clustering and data visualization. Data were collected from the most used dataset published by Johns Hopkins University.

The experimental results are still in progress and future works will present the final version of the process which will cluster more than 200 countries into dynamic communities according to the similar spreading of Covid-19. Also, we study the possibility to integrate graph neural networks to our SC-based approach in order to have deep learning model.

References

[1] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, *et al.*, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery,* 2020.

[2] A. Mollalo, B. Vahedi, and K. M. Rivera, "GIS-based spatial modeling of COVID-19 incidence rate in the continental United States," *Science of The Total Environment,* vol. 728, p. 138884, 2020.

[3] P. Liu and X.-z. Tan, "2019 novel coronavirus (2019-nCoV) pneumonia," *Radiology,* vol. 295, pp. 19-19, 2020.

[4] R. Moulay Taj, Z. Ait El Mouden, A. Jakimi, and M. Hajar, "Towards Using Recurrent Neural Network for Predicting Influenza-Like Illness: Case Study of Covid-19 in Morocco," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 9, No 5, 2020.

[5] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871,* 2020.

[6] K. Elasnaoui and Y. Chawki, "Using X-ray Images and Deep Learning for Automated Detection of Coronavirus Disease," *Journal of Biomolecular Structure and Dynamics,* pp. 1-22, 2020.

[7] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055,* 2020.

[8] C.-T. Yang, Y.-A. Chen, Y.-W. Chan, C.-L. Lee, Y.-T. Tsan, W.-C. Chan, *et al.*, "Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources," *The Journal of Supercomputing,* pp. 1-27, 2020.

[9] Z. A. El Mouden, A. Jakimi, and M. Hajar, "An application of spectral clustering approach to detect communities in data modeled by graphs," in *Proceedings of the 2nd International Conference on Networking, Information Systems & Security*, ACM International Conference Proceeding Series, Article No 4, 2019.

[10] Z. A. El Mouden, R. M. Taj, A. Jakimi, and M. Hajar, "Towards for Using Spectral Clustering in Graph Mining," in *International Conference on Big Data, Cloud and Applications*, *Communications in Computer and Information Science,* vol. 872, 2018 pp. 144-159.

[11] M. Afzalan and F. Jazizadeh, "An automated spectral clustering for multi-scale data," *Neurocomputing,* vol. 347, pp. 94-108, 2019.

[12] W. Casaca, G. Taubin, and L. G. Nonato, "Graph laplacian for spectral clustering and seeded image segmentation," in *Anais do XXVIII Concurso de Teses e Dissertações*, 2020, pp. 31-36.

[13] D. R. DeFord and S. D. Pauls, "Spectral clustering methods for multiplex networks," *Physica A: Statistical Mechanics and its Applications,* vol. 533, p. 121949, 2019.

[14] D. Das, K. Santosh, and U. Pal, "Truncated inception net: COVID-19 outbreak screening using chest X-rays," *Physical and engineering sciences in medicine,* pp. 1-11, 2020.

[15] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," *Chaos, Solitons & Fractals,* vol. 140, p. 110122, 2020.

[16] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arXiv preprint arXiv:2003.11597,* 2020.

[17] K. Rajagopal, N. Hasanzadeh, F. Parastesh, I. I. Hamarash, S. Jafari, and I. Hussain, "A fractional-order model for the novel coronavirus (COVID-19) outbreak," *Nonlinear Dynamics,* vol. 101, pp. 711-718, 2020.

[18] R. Dandekar and G. Barbastathis, "Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning," *medRxiv,* 2020.

[19] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," *arXiv preprint arXiv:2001.08317,* 2020.

[20] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa, "Spatial analysis and GIS in the study of COVID-19. A review," *Science of The Total Environment,* p. 140033, 2020.

[21] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing,* vol. 17, pp. 395-416, 2007.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, pp. 888-905, 2000.

[23] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics,* vol. 39, pp. 1878-1915, 2011.

[24] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels, "Scalable community detection with the louvain algorithm," in *2015 IEEE*

International Parallel and Distributed Processing Symposium, 2015, pp. 28-37.

[25] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks,* vol. 20, pp. 61-80, 2008.

[26] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang*, et al.*, "Graph neural networks: A review of methods and applications," *arXiv preprint arXiv:1812.08434,* 2018.

[27] S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," *Journal of medical virology,* vol. 92, pp. 632-638, 2020.

[28] Z. Ait El Mouden and A. Jakimi, "k-eNSC: k-estimation for Normalized Spectral Clustering," 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), IEEE, 2020.