Research Article

Networks of Countries Defined by the Dynamics of the COVID-19 Pandemic

Dmytro Lande [a], *Leonard Strashnoy* [b] (✉)

[a] *Institute for Information Recording of NAS of Ukraine*
[b] *UCLA, Institute Infectious disease department, USA*

ABSTRACT:

A technique for forming, clustering and visualizing so-called correlation networks of countries determined by the dynamics of the COVID-19 pandemic is here proposed. The links between countries as nodes of such networks correspond to the values of correlations between sets of parameters corresponding to the dynamics of the pandemic in these countries. To build network structures for each node (country), vectors are formed - arrays of numbers corresponding to the dynamics of the pandemic (in one case - the dynamics of daily mortality, in the second - the dynamics of infection). For this purpose, data obtained from an external source - an aggregator of such data - is used. This approach, in contrast to the existing ones, has such advantages as a relatively low dimension of vectors-parameters corresponding to countries; a reliable mathematical basis for correlation analysis; objectivity - for the "purity" of data corresponds to a reliable data aggregate; the use of standard software tools; and the relative ease of implementation. This method can be used in analytical systems for various purposes to analyze arrays of entities without explicit relationships between them. Correlation networks can be considered as the basis for constructing probabilistic networks and applying fuzzy semantic network technologies for further analysis with the use of experts, and decision support systems.

KEYWORDS:

social media monitoring, cyber security, OSINT, Big Data, Cyber Aggregator
Correlation Network, Pandemic Dynamics, Data Aggregation, Network visualization, Cluster Analysis, Modularity

✉ Corresponding Author E-mail: lstrashnoy@gmail.com,

Introduction

Information networks can be considered as a basis for conducting cluster analysis - identifying groups of similar documents and visualizing them. However, these network structures are not always explicitly defined. Of course, if we are talking about explicit networks, nodes and connections between them, then there are no special problems. But how do you build a network to apply a wide range of methods and tools for processing it, to get and interpret the results, if the researcher has only some node-entities, for example, countries, but no defined edges - as connections between them? If each object in the system can be represented as a uniform multidimensional vector of parameters, you can use classification or cluster analysis methods to identify groups of similar documents. In many well-known models of information search for a document or an array of documents that correspond to a specific topic, the vector of the weight of words included in it is matched in particular.[1] In this case, there are several metrics for determining the distance between documents, the most prominent of which is Euclidean.

Each entity in s_k in the set $S = \{s_k\}_{k=1}^{|S|}$ is matched with a vector of parameter values $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, where $n = |G|$ is the number of elements in the set of parameters. The correlation between entities s_j (a_{ij}) can be defined, for example, as the cosine of the angle between the corresponding vectors $\overline{w^i}$ and $\overline{w^j}$:

$$a_{ij} = \frac{\left(\overline{w^i}, \overline{w^j}\right)}{\left|\overline{w^i}\right|\left|\overline{w^j}\right|} = \frac{\sum_{k=1}^{n} w_k^i w_k^j}{\sqrt{\sum_{k=1}^{n}\left(w_k^i\right)^2}\sqrt{\sum_{k=1}^{n}\left(w_k^j\right)^2}}, \quad (1)$$

where a_{ij} are elements of the square adjacency matrix $A = \left\|a_{ij}\right\|$.

In this paper, a method is proposed that also matches the entity-country vector (the so called vector of the dynamics of the pandemic process), which corresponds to the processes of infection or mortality over time. More specifically, each era is assigned a number – the number of cases of Coronavirus infection or deaths from it in the corresponding country. The dimension of this vector corresponds to the number of days and the length of the time interval during which the observations were made.

Goal

The purpose of this paper is to present a methodology for forming, clustering, and ranking nodes and visualizing so-called correlation networks, structure graphs, and links between nodes (entities, countries) that correspond to the values of correlations between sets of parameters corresponding to these entities. In this case, the values of time sequences determined by the dynamics of the pandemic for each individual country are considered as sets of parameters.

At the same time, it should be noted that correlation does not directly mean causal relationships, so correlation networks cannot be considered as causal, semantic maps. At the same time, correlation, along with other criteria, can be considered as the basis for probabilistic estimates. In other words, correlation networks can be considered as the basis for constructing probabilistic networks, as the basis for applying fuzzy semantic network technologies for further expert analysis.

To build network structures for each entity/country, vectors are formed that correspond to the dynamics of the pandemic in that country. For this purpose, it is planned to use a system of aggregation of news on the pandemic in different countries of the world, such as the World Health Organization (WHO) or Our World In Data (a project of the Global Change Data Lab).[2]

Such systems allow obtaining arrays of numbers corresponding to the processes of the COVID-19 pandemic in various countries. To get these arrays visually, you can log in to the corresponding web sites via the interface.

After the formation of vectors corresponding to individual countries, a correlation network is formed, which can be considered as a storage and visualization of nodes-countries that are objectively related to each other. [3]

Indeed, it is possible to form vectors of pandemic dynamics for different countries, the relationship between which is not always clear.

Sources of Information and Timelines

This work uses data from the World Health Organization (WHO) as a source of information https://www.who.int/emergencies/diseases/novel-coronavirus-2019/ situation- reports/ and in Our World In Data. Within this source (https://ourworldindata.org/coronavirus-source-data) data is daily renewed and presented in a “cleaned up” format, without the anomalous spikes that have been connected with technical failures, with a high level of integration, in the formats XLSX (https://covid.ourworldindata.org/data/owid-covid-data.xlsx) CSV (https://covid.ourworldindata.org/data/owid-covid-data.csv), JSON (https://covid.ourworldindata.org/data/owid-covid-data.json) (Fig. 1).

Addresses of individual data arrays:

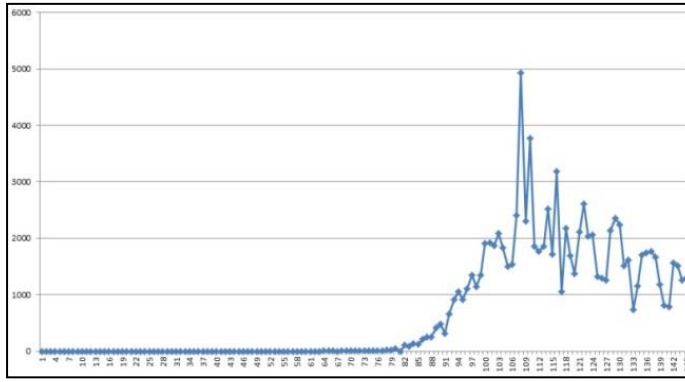
- Total confirmed cases:
  https://covid.ourworldindata.org/data/ecdc/total_cases.csv
- Total deaths:
  https://covid.ourworldindata.org/data/ecdc/total_deaths.csv
- New confirmed cases:
  https://covid.ourworldindata.org/data/ecdc/new_cases.csv
- New deaths:
  https://covid.ourworldindata.org/data/ecdc/new_deaths.csv
- All four metrics:
  https://covid.ourworldindata.org/data/ecdc/full_data.csv
- Population data:
  https://covid.ourworldindata.org/data/ecdc/locations.csv

Figure 1: A fragment of the interface on the website ourworldindata.org

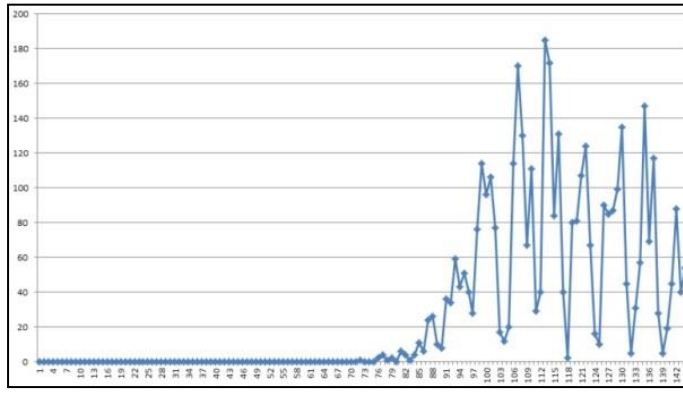
In Fig. 2. the Daily Mortality Rate is presented (a) throughout the USA, and (b) Sweden from the data set brought over from the website ourworldindata.org. starting 01/01/2020.

Method

Below, a method is proposed for building a network of interconnections of entities (countries), which consists of the following stages.



a)



b)

Figure 2: Daily Mortality Rate ourworldindata.org a) USA; b) Sweden (on the horizontal axis - date, on the vertical axis - number of deaths)

1) For each country from the data file downloaded from the aggregator, vectors of the dynamics of the pandemic process for a certain period are determined (selected from 15.03.2020 to 30.06.2020), corresponding to the severity or infectability, similar to those shown in Figure 2.
2) The set of maximal cross-correlations between the obtained vectors is calculated, and the corresponding correlation matrix is formed with the elements in the notation of the formula (1):

$$a_{ij}(m) = \max_{m} \frac{\sum_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum_{k=1}^{n-m} \left(w_{k}^{j}\right)^{2}}}. \quad (2)$$

The max function is used for the reasons that processes that are similar in nature may have similar dynamic behavior, but possibly with a time shift.

3) The adjacency matrix is formed in accordance with formula (2) and stored in a CSV file (Figure 3). Due to the fact that there are connections between all nodes in the adjacency table, according to [4], connections whose value is less than a certain threshold are ignored. The choice of this threshold completely depends on the experience of analysts. In the described information technology the formed matrix is transmitted for processing and visualization to the network structure analysis system Gephi (https://gephi.org/). Gephi is a widely used program for visualization and analysis of network structures, provides fast modularization, efficient data research, as well as visualization of large-scale networks. At the same time, the CSV adjacency matrix for the Gephi system has some features that need to be taken into account (zeros on the diagonal, the location of the characters ";") etc.

A1

	A	B	C	D	E	F	G	H	I	J	K
1		AFG	DZA	AUS	AUT	BGD	BEL	BRA	CAN	CHL	CHN
2	AFG	0.000	0.034	0.000	0.000	0.716	0.000	0.671	0.227	0.645	0.046
3	DZA	0.034	0.000	0.487	0.440	0.060	0.547	0.000	0.322	0.064	0.517
4	AUS	0.000	0.487	0.000	0.670	0.000	0.698	0.000	0.453	0.000	0.407
5	AUT	0.000	0.440	0.670	0.000	0.000	0.796	0.000	0.518	0.000	0.303
6	BGD	0.716	0.060	0.000	0.000	0.000	0.000	0.811	0.278	0.685	0.000
7	BEL	0.000	0.547	0.698	0.796	0.000	0.000	0.000	0.644	0.000	0.251
8	BRA	0.671	0.000	0.000	0.000	0.811	0.000	0.000	0.535	0.621	0.000
9	CAN	0.227	0.322	0.453	0.518	0.278	0.644	0.535	0.000	0.122	0.184
10	CHL	0.645	0.064	0.000	0.000	0.685	0.000	0.621	0.122	0.000	0.000
11	CHN	0.046	0.517	0.407	0.303	0.000	0.251	0.000	0.184	0.000	0.000
12	COL	0.695	0.122	0.000	0.000	0.819	0.000	0.658	0.069	0.633	0.000
13	DNK	0.000	0.470	0.693	0.744	0.000	0.865	0.000	0.587	0.000	0.280

corr_d

Figure 3: Displaying the example adjacency matrix in Excel. Country names in accordance with [5]

4) This matrix is uploaded in CSV format to the Gephi system. This system has a number of modes, among which the "data Lab" mode is used for monitoring network characteristics. In this mode, in addition to the usual degrees of matrix nodes, you can calculate their values by PageRank, Hits, modularity, and so on. In addition, there are options for ranking matrix nodes (entities) by these parameters (Figure 4).

Id	Modularity Class
HUN	7
IRL	7
JPN	7
PRT	7
TUR	7
AUS	7
AUT	7
BEL	7
DNK	7
FRA	7
DEU	7
GRC	7
ITA	7
NLD	7
NOR	7
ESP	7
CHE	7
ECU	6
ISR	6
CHN	6
IRN	5
KOR	5
BGD	4
BRA	4

Figure 4: Fragment of the table in the "data Lab" mode of the Gephi system

5) Object group modularity classes are defined and the loaded network structure is then cauterized [5]. Modularity is calculated as the difference between the fraction of edges within a cluster in the network under consideration and the expected fraction of edges within a cluster in a network where vertices have the same degree as in the original one, but the edges are randomly distributed. The modularity of the network can be expressed by the formula:

$$Q = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - \frac{k_i k_j}{2m}\right]\ \delta(c_i, c_j), \tag{3}$$

where a_{ij} is an element of the A adjacency matrix, m – is the number of edges in the graph, k_i, k_j are the degrees of nodes *i* and *j* respectively, and δ – is the Kronecker delta function (shows whether nodes are located *i* and *j* in the same module).

6) Network visualization is performed in the Gephi system. Results of the entity network (country) initialization are shown in Fig. 5 and 6.
7) At the last stage, an expert interpretation of the results takes place. Data in tabular and graphical form is transmitted to experts (virologist or epidemiologists, in our case).

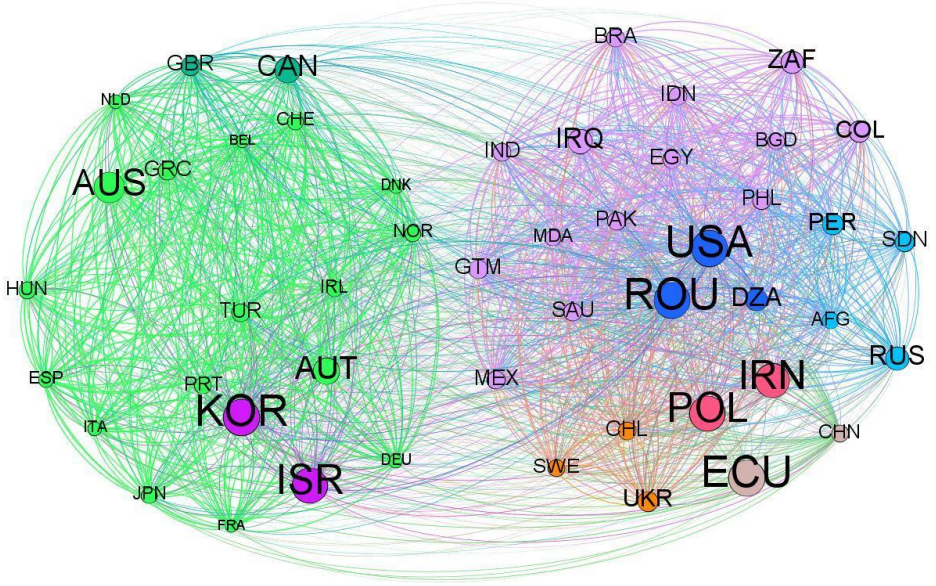


Figure 5: Network of entities (countries) in the Gephi environment, built on the basis of correlations of infection diagnoses per day

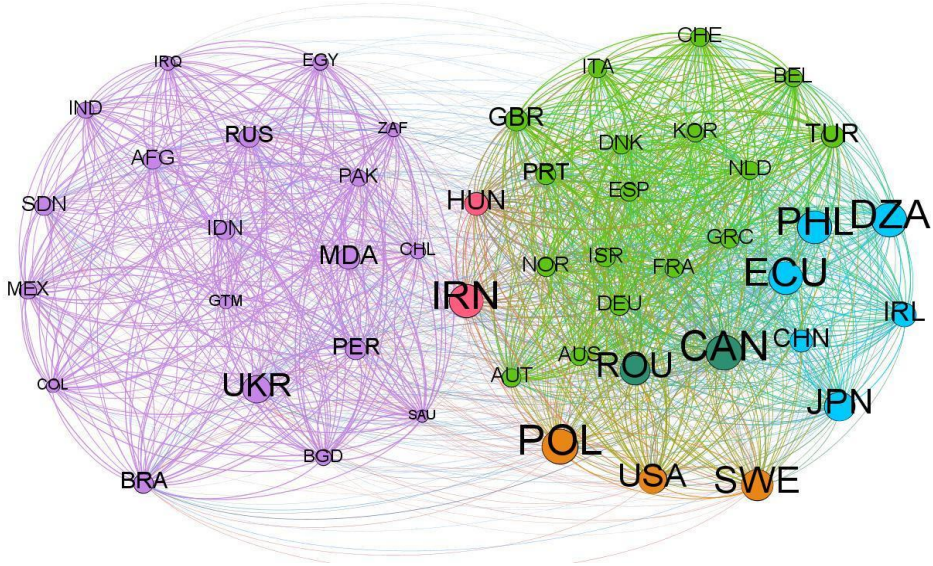


Figure 6: Network of entities (countries) in the Gephi environment, built on the basis of correlations of daily mortality

In this example, 8 clusters are identified, but it is logical to group them accordingly:

1. Developed countries
2. Countries of the "third world"
3. Countries with "multi-wave" dynamics, such as the United States, Sweden, Iran, Israel.

In the example shown in Fig. 5 and 6 of networks corresponding to the actual processes of the COVID-19 pandemic, we can note their great similarity when divided into clusters.

To substantiate this fact, a matrix corresponding to the distribution of objects in clusters is considered. The columns of this matrix correspond to objects (in our case, countries), and the rows correspond to clusters. In this matrix, the element value is 1 if the object belongs to the corresponding cluster and 0 if it does not. To determine the similarity level of two such matrices *A* and *B* when allocating *N* objects to *M* clusters the following formula is proposed:

$$\chi = \max_{\Omega} \frac{\sum_{i=1}^{N}\sum_{k=1}^{M} a_{ik} b_{ik}}{N}, \tag{4}$$

where Ω is the set of all permutations of matrix rows *A*, a_{ik} and b_{ik} – elements of matrices *A* and *B*.

In the example discussed above, the value χ exceeds 0.7, although if you randomly distribute objects across 3 clusters, this value would not increase by an average of 1/3.

Conclusions

This paper describes the concept of a correlation network, provides a methodology for its formation, clustering, ranking and visualization of nodes that correspond to the vectors of the pandemic dynamics.

This approach, unlike the existing ones, has the following advantages:

- the relatively low dimension of the vectors of parameters corresponding to countries;
- reliable mathematical basis for correlation analysis;
- objectivity – a reliable data aggregator is responsible for the "purity" of data;
- use of standard software tools;
- relative ease of implementation (ready-made software systems such as Gephi, Matlab, Excel, the R language, etc. can be used).

The presented method also can be used in analytical systems of various purposes for the analysis of arrays of entities without clearly defined connections between them.

Examples of entities for which the developed method can be applied, other than in the case discussed, are the following:

- handlings the spread of diseases;
- political leadership characterized by an attitude to various spheres of public life;
- consumers of products – here parameters could be sellers or the sources of products;
- mass media as content entities, in this case parameters can be words as indicators of "fakes" in the headings of articles that are published in these publications.

The found patterns can be passed on to the experts (in this case, epidemiologists) to support decision-making.

Potential future enhancements:

- Volumes of COVID diagnosed and death cases as well as its derivatives (such as percentages of the population) could be used as the model parameters;
- Time lag between epidemics starting in specific countries would be accommodated in the Gephi model.

We would like to thank Professor Andrew Snarsky of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute" for his recommendations and discussion of the results of this article.

References

1 Christopher D. Manning. An Introduction to Information Retrieval. Cambridge University Press, 2009. 569 p.

2 Our World In Data Project. URL: https://ourworldindata.org/coronavirus-source-data

3 John W. Foreman. Using Data Science to Transform Information into Insight Data Smart. Wiley, 2013.

4 Ken Cherven. Mastering Gephi Network Visualization. Packt Publishing, 2015.

5 ISO 3166-1 alpha-3. URL: https://www.iso.org/iso-3166-country-codes.html