Analysis of Worldwide Time-Series Data Reveals Some Universal Patterns of Evolution of the SARS-CoV-2 Pandemic

Marija Mitrović Dankulov¹, Bosiljka Tadić²,³ and Roderick Melnik⁴,⁵

¹Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia, ²Department for Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia, ³Complexity Science Hub, Vienna, Austria, ⁴MS2Discovery Interdisciplinary Research Institute, M2NeT Laboratory and Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada, ⁵BCAM–Basque Center for Applied Mathematics, Bilbao, Spain

Predicting the evolution of the current epidemic depends significantly on understanding the nature of the underlying stochastic processes. To unravel the global features of these processes, we analyse the world data of SARS-CoV-2 infection events, scrutinising two 8-month periods associated with the epidemic’s outbreak and initial immunisation phase. Based on the correlation-network mapping, K-means clustering, and multifractal time series analysis, our results reveal several universal patterns of infection dynamics, suggesting potential predominant drivers of the pandemic. More precisely, the Laplacian eigenvectors localisation has revealed robust communities of different countries and regions that break into clusters according to similar profiles of infection fluctuations. Apart from quantitative measures, the immunisation phase differs significantly from the epidemic outbreak by the countries and regions constituting each cluster. While the similarity grouping possesses some regional components, the appearance of large clusters spanning different geographic locations is persevering. Furthermore, characteristic cyclic trends are related to these clusters; they dominate large temporal fluctuations of infection evolution, which are prominent in the immunisation phase. Meanwhile, persistent fluctuations around the local trend occur in intervals smaller than 14 days. These results provide a basis for further research into the interplay between biological and social factors as the primary cause of infection cycles and a better understanding of the impact of socio-economical and environmental factors at different phases of the pandemic.

Keywords: complex networks, k-means, time-series analysis, spectral analysis, community structure, SARS-CoV-2

1 INTRODUCTION

In cooperative social dynamics [1, 2], the genesis of a collective phenomenon arising from contagious social interactions involves mechanisms of self-organised criticality [3, 4]. It depends on each individual involved, based on its actual contacts, psychology and behaviour. In the presence of viruses, these mechanisms are additionally shaped by firm biological factors. Recent developments of SARS-CoV-2 pandemic [5, 6] revealed a specific global phenomenon emerging from the stochastic multi-scale processes. The infection incidence occurs with a high temporal resolution at the
interactions between the virus and human hosts, whose biological features, social behaviours and mobility \cite{7} significantly contribute to the epidemic’s spreading \cite{8}. At the molecular scale, the virus-host interactions \cite{9–11} crucially depend on the virus biology and genetic factors determining the host’s immunity towards the virus in question \cite{12, 13}. Thus, the occurrence of an infection event and the infection manifestation may lead to a range of different scenarios from asymptomatically infected to severe health issues and fatalities \cite{14–17}. Multiple other factors may play a role \cite{18}, depending on the population genetic features and social life \cite{19}. They include cultural, political and economic aspects, official and spontaneous reaction to the crisis, and the organisation of the health care system, all of which may significantly differ between different geographical locations \cite{20}. Moreover, the actual impact of these factors changes over time as the epidemic develops, in particular, since the appropriate vaccines targeting SARS-CoV-2 are available, thus enabling potentially substantial changes due to massive immunisation of the population given the theoretical analysis in \cite{22–24}. Attempts were made to identify different parameters that may influence the epidemic and estimate their mutual interdependence and impact. For example, the human-development index, built-up-area-per-capita, and the immunisation coverage appear among the statistically high-ranking drivers of SARS-CoV-2 epidemic \cite{18}.

In addition, temporal variations occur at all scales, from the virus mutations \cite{11} to changed behaviours of each individual and population groups, e.g., due to the government imposed measures \cite{6, 25}, or adaptation caused by the awareness of the current epidemiological situation \cite{26, 27}. These variations increase the stochasticity of the infection and contact processes, making the prediction of their output even more difficult. For real-time epidemic management and the predictions of further developments, it is crucial to understand the nature of the underlying stochastic processes and the factors that can significantly influence them. For this purpose, the empirical data analysis and theoretical modelling \cite{28} provide complementary views of these complex processes. For example, agent-based models capture the interplay of the bio-social factors at the elementary scale of the virus-host interactions at high temporal resolution \cite{8, 29–37}. On the other hand, more traditional compartmental models \cite{38} consider a coarse-grained picture of the population groups having different roles in the process. Another research line aims at the mathematical description of the exact empirical data, in particular, for the outbreak phase \cite{39, 40}. For instance, different studies provided tangible arguments for the cause of the changing shape of the infection curve comprising the appearance of linear and power-law segments \cite{41, 42}, prolonged stagnation periods, and multiple waves \cite{43}. Since the beginning of the epidemic, empirical data were collected over different countries or provinces \cite{44}. Despite the coarse-grained spatial and temporal structure (daily resolution), these data may contain relevant information about the temporal aspects of the epidemic at different geographical locations. Previous studies, based on the empirical data regarding the dynamics of interacting units in many complex systems, provided valuable information about the related stochastic processes. Some striking examples across different spatial and temporal scales include the influence of the world financial index dynamics on different countries \cite{45, 46}, traffic jamming \cite{47, 48}, brain-to-brain coordination dynamics \cite{49, 50}, and the cooperative gene expressions along different phases of the cell cycle \cite{51, 52}. Similarly, the collected data of SARS-CoV-2 spreading enable a possibility to investigate the infection dynamics in various details and more appropriate modelling of the emergent behaviours. In this respect, a larger-scale picture may emerge by studying temporal fluctuations of the world infection dynamics. More subtle questions regard the indicators for hidden mechanisms arising from the interplay of the above-mentioned biological factors and different social behaviours \cite{8, 27, 29, 53, 54} behind the observed epidemic development.

In this work, we address some of these critical issues aiming to unveil the inherent features of infection dynamics by studying time-series data that are publicly available at GitHub \cite{44} collected over different countries or regions (provinces). Using the datasets of the daily recorded number of confirmed infection cases, we consider two separate segments of time series. Defining two distinct 8-month periods in the epidemic’s evolution is motivated by the appearance of SARS-CoV-2 vaccines in the latter period, enabling pharmaceutical intervention measures not available in the outbreak phase, cf. Figure 1. Namely, the records for the first 8 months of the epidemic, starting from the first registered case in each country, represent the epidemic’s outbreak phase. Meanwhile, the last 8 months (preceding the data collection on 30 September 2021), during which the pharmaceutical intervention was available in most of the countries, characterises the initial immunisation phase of this pandemic. Our quantitative analysis comprises three levels of information: the network mapping and spectral analysis, K-means clustering of pairs of time series, and detrended fractal analysis of individual time series. Each of these methods provides just partial information about the studied dynamics. We combine them to create a comprehensive picture of the course of the epidemic in different countries and how they relate to each other. In addition to quantifying the differences between the outbreak and immunisation phase, our results reveal two global features of the SARS-CoV-2 pandemic. Firstly, the worldwide groups of countries (and provinces) robustly appear in clusters having a similar temporal evolution of the infection dynamics. This clustering suggests that the environmental and socio-economic factors and government-imposed measures can certainly influence small-scale fluctuation characteristics of the clusters but do not significantly change the course of the process on larger scales. Secondly, the epidemic evolution exhibits ubiquitous waves driven by the cyclic infection dynamics, where several typical cycles appear associated with the identified clusters. Again, the shape of these specific cycles coincides with the mentioned clustering mechanisms. Hence, their origin and potential control will remain challenging within purely social measures. A more detailed analysis of the complex feedback between biological and social factors at all scales is needed.
2 MATERIALS AND METHODS

2.1 Data Acquisition, Preparation, and Mapping

We consider the worldwide data of the number of new infection cases downloaded from GitHub [44]. The dataset contains the number of daily detected new cases for 279 countries including separated data for some provinces. For this work, we select time series in two eight-month periods comprising the epidemic’s outbreak phase (starting from the first registered case in a given country or province) and the immunisation phase (22 January 2020 until 30 September 2021). The corresponding number of countries and provinces with the active epidemic’s data traced in both periods is 255. For instance, the first case in France was detected on 24 January 2020, and thus the outbreak time series covers the period from that date until 19 September 2020. However, Slovenia had the first registered case on 5 March 2020; hence its outbreak time series cover 5 March until 30 October 2020. Meanwhile, the immunisation period is from 3 February 2021 to 30 September 2021, equal for all considered countries and provinces.

By mapping these datasets, we obtain two correlation networks for the outbreak and immunisation phase, respectively, where the network’s links stand for significant positive correlations. We first compute the Pearson’s correlation coefficient for the corresponding pairs \((i, j)\) of the time series

\[
C_{ij}^\tau = \frac{1}{N_T - 1} \sum_{t=1}^{N_T} \frac{X_i^\tau(t) - \mu_i^\tau}{\sigma_i^\tau} \frac{X_j^\tau(t) - \mu_j^\tau}{\sigma_j^\tau},
\]

where \(\tau \in \{O, V\}, \mu_i^\tau\) is average value of the time series of country \(i\) during period \(\tau\), \(\sigma_i^\tau\) is standard deviation of time series \(X_i^\tau(t)\), and \(N_T = 240\) is the length of time series. To remove spurious correlations, we apply the filtering procedure standardly used in these type of network mapping [47, 49, 51]. More precisely, the matrix elements \(C_{ij}^\tau\) are first transformed to the interval \([0, one]\) by \(CP_{ij}^\tau = \frac{1}{2} (C_{ij}^\tau + 1)\), and then multiplied by a factor \(M_{ij}\) which is obtained in the following way. From the rows \(i\) and \(j\), the diagonal elements are removed and the considered elements \(CP_{ij}^\tau\) and \(CP_{ji}^\tau\) are placed at the beginning of the row \(i\) and \(j\), respectively, thus obtaining two \(n = N - 1\) dimensional vectors \(CP_i^\tau\) and \(CP_j^\tau\). Then \(M_{ij}\) is computed as Pearson’s coefficient between these two vectors. The matrix element of the filtered correlation matrix \(\tilde{C}_{ij}^\tau = M_{ij}^\tau CP_{ij}^\tau\) is then mapped back to the interval \([-1, 1]\).

Finally, the elements of the network’s adjacency matrix are defined as \(A_{ij}^\theta = 1\) when the matrix elements \(\tilde{C}_{ij}^\tau > \theta\) exceed a specified threshold value \(\theta\), and zero otherwise. The threshold value \(\theta\) is determined concerning the network’s spectral properties, as described below.

2.2 Network’s Spectral Analysis and Community Detection

The above-described data mapping should lead to undirected unweighted networks; the nodes represent countries (or provinces), and links indicate the positive correlations between infection incidences exceeding a threshold \(\theta\). We use the spectral properties of networks to obtain the adequate threshold value, where the guiding criteriums are the network’s sparseness and the relative stability of the community structure. Starting from \(\theta = 0\), we increase it by the value 0.05 and solve the eigenvalue problem of the normalised Laplacian related to the network obtained for considered \(\theta > 0\) using Kolmogorov-Smirnov (KS) distance. For each \(\theta > 0\) we obtain one KS-distance and plot its dependence of \(\theta\), see Figure 2B. The KS-distance has a minimum of around \(\theta = 0.5\) for the outbreak and immunisation phase. We use this value of \(\theta\) to obtain the networks used in our analysis.

We study the community structure of the networks for the outbreak and immunisation period using spectral analysis and the eigenvalue problem of the normalised Laplacian related to the network’s adjacency matrix. In mathematics theory [55, 56], the number of smallest non-zero eigenvalues of the Laplacian matrix is a good indicator of the number of communities. The matrix elements of the normalised Laplacian for undirected binary network represented by the adjacency matrix \(A\) are defined as...
where \( q_i \) and \( q_j \) are degrees of nodes \( i \) and \( j \). For the normalised Laplacian \([2]\), we solve the eigenvalue equation \( \mathbf{L} \mathbf{v} = \lambda \mathbf{v} \) and determine all eigenvalues and eigenvectors. In the case of a connected network, these eigenvalues are non-negative. One zero-eigenvalue appears with strictly positive eigenvector’s components \([55]\). Consequently, the orthogonal eigenvectors corresponding to the three smallest non-zero eigenvalues localise on the communities of the network. Hence, the scatter plot of the components of these eigenvectors shows a branching structure. Each branch contains indexes of the non-zero eigenvector components, that is, the nodes belonging to a network’s community \([56]\). The size of the q-core of the networks is determined by removing the nodes with the increasing degree \( q \). Several other graph properties are determined, and the networks are visualised using Gephi software \([57]\).

### 2.3 K-Means Clustering of Time Series

The implementation of the K-means algorithm for clustering of time series in Python known as `tslearn` \([58]\) is used. K-means is an unsupervised machine learning algorithm that aggregates data points according to similarities, starting with \( K \) randomly positioned centroids. Based on these centroids, data points are assigned to the centroid closest to that data point according to some distance metric. The algorithm consists of a certain number of iterative (repetitive) calculations used to optimise the positions of the centroids. Considering each time series of length \( N_f \), as a data point in \( N_f \) dimensional space, the appropriate measures enable calculating the distances between these data points. We use the Dynamic Time Wrapping (DTW) algorithm to align time series with centroids and measure their similarities. The DTW is a widely used algorithm measuring similarities between time series and their classification. It does not transform the time series; it only finds the minimal distance between time series beyond simple correlation. Specifically, it performs an optimal alignment between two time series by matching the indices from the first time series to the second time series, subject to several constraints. The mapping of indices from the first series to the second series must be monotonically increasing. For the indices \( i > j \) from the first time series, there must be two indices from the second series \( l > k \) such that \( i \) is matched with \( l \) and \( j \) is matched with \( k \). Meanwhile, the first index from the first series must match the first index of the second time series, and similarly, the last index from the first series must be matched to the last index of the second time series, but these points may have more other matches. The optimal alignment is the one that satisfies all of these restrictions with the minimal cost, where cost is the sum of absolute differences of values for each matched pair of indices. The DTW distance in the K-means algorithm is the value of cost. We use the K-means algorithm with DTW distance to cluster time series and find centroids. Each centroid is again a time series that describes the average behaviour of the time series belonging to one cluster.

#### 2.4 Trends and Fractal Analysis of Time Series

Temporal fluctuations are studied by the fractal detrended analysis of each time series. For each time series \( x(k), k = 1, 2, \cdots, T \), the profile \( Y(i) = \sum_{k=1}^{N_f}(x(k) - \langle x \rangle) \) of the time series is divided in \( N_s \) segments of the length \( n \). The fluctuation function \( F_q(n) \) with the varied segment length \( n \) is defined as

\[
F_q(n) = \left( \frac{1}{N_s} \sum_{i=1}^{N_s} \left[ F^2(\mu, n) \right]^{q/2} \right)^{1/q}, \quad n^h
\]

Here, \( F^2(\mu, n) = \sum_{i=1}^{n} [Y((\mu - 1)n + i) - Y_\mu(i)]^2 \) is the standard deviation from a local trend \( y_\mu(i) \) on the segment \( \mu \). For \( q = 2 \), we determine the Hurst exponent \( h_2 \) from the straight-line segments of the log-log plot of the fluctuation function \( F_2(n) \). For the multifractal analysis, the values of \( q \in [-4, 4] \) are varied.

To determine cyclic trends, we use the local adaptive detrending algorithm, see \([59, 60]\), where time series is divided into segments of the length \( 2m + 1 \) overlapping over \( m + 1 \) points. The polynomial interpolation is applied in each segment, and its contribution in the overlapped region is weighted such that it decreases linearly with the distance from the segment’s centre. As stated in the Introduction, we consider worldwide recorded time series of the infection cases. For illustration, a few examples of time series recorded in different countries are shown in Figure 1.
probability distributions of $fi$ have a peak at a value the outbreak and immunisation period. While both probability values of correlation coefficients differ drastically from the corresponding outbreak and immunisation period networks at $\theta = 0$ containing all positive correlations. The resulting networks for $\theta = 0.5$ are visualised in Figure 3. See also Supplementary Information (SI) for more details.

The giant connected component of each network exhibits a community structure, i.e., the occurrence of groups of nodes that are better connected among themselves than with the nodes outside that group, cf. Figure 3. The identity of nodes comprising each community is determined using the localisation of the eigenvectors associated with the three lowest nonzero eigenvalues of the normalised Laplacian operator [55], as explained in Methods [56]. The eigenvalues of the normalised Laplacians for two networks are shown in ranking order in Figure 4, middle panel. Several lowest nonzero eigenvalues appear to be separated from the bulk in both networks. This network feature is compatible with the existence of mesoscale communities, on which the corresponding eigenvectors tend to localise [55, 56]. The scatterplots of the eigenvectors associated with three lowest nonzero eigenvalues, see Figure 4, show three differentiable branches, here indicated as G1, G2, G3 for the outbreak, and g1, g2, g3 for the immunisation phase network. The indexes with a nonzero component of the eigenvectors in each branch mark the IDs of the nodes belonging to the corresponding community. The complete lists of nodes in each community (group) are given in Supplementary Tables S1–S6 in SI.

Even though both networks exhibit three major communities, the structural differences between the two networks in Figure 3 are apparent. They indicate the corresponding differences in the fluctuations of the infection rates in the world regions during the immunisation phase, compared to the epidemic’s outbreak, when

---

**2.5 The Correlation Networks Mapping in the Outbreak and Immunisation Phase**

The network mapping is based on the cross-correlation coefficient $C_{ij}$ of the pairs of time series $\{i, j\}$ and a suitably selected threshold. Hence, the correlations exceeding the threshold $\theta$ are accepted, making the adjacency-matrix elements $A_{ij}(\theta) = \Theta(C_{ij} - \theta) - \delta_{ij}$ of an undirected unweighted network. Before selecting the threshold, a filtering procedure was applied to the complete correlation matrix to enhance the positive correlations of interest in this work (see Methods). The applied methodology was proved useful in quantifying correlations of interest in this work (see Methods). The applied methodology was proved useful in quantifying correlations of interest in this work (see Methods).

After selecting the threshold, the filtering procedure was applied to the complete correlation matrix to enhance the positive correlations of interest in this work. The applied methodology was proved useful in quantifying correlations of interest in this work.

The network mapping is based on the cross-correlation coefficient $C_{ij}$ of the pairs of time series $\{i, j\}$ and a suitably selected threshold. Hence, the correlations exceeding the threshold $\theta$ are accepted, making the adjacency-matrix elements $A_{ij}(\theta) = \Theta(C_{ij} - \theta) - \delta_{ij}$ of an undirected unweighted network. Before selecting the threshold, a filtering procedure was applied to the complete correlation matrix to enhance the positive correlations of interest in this work. The applied methodology was proved useful in quantifying correlations of interest in this work (see Methods). The applied methodology was proved useful in quantifying correlations of interest in this work (see Methods).

**FIGURE 3** Giant connected components of the correlation networks at the threshold $\theta = 0.5$ for the outbreak period (A) and the immunisation period (B). Red, green and blue colours indicate groups of nodes in three respective communities G1, G2, G3 in the outbreak network, and g1, g2, g3 in the immunisation period network, determined by the eigenvector-localisation, see text and Figure 4. Unclassified borderline nodes are shown in white colour. Labels on nodes identify the corresponding country or province. The complete lists of nodes in each community are given in Supplementary Tables S1–S6 in Supplementary Information.
the whole population was practically susceptible to the infection. These differences are quantified by several graph measures, see Figures 2A–D and Table 1. Compatible with these graph-theory measures are the span of the exponentially-decaying degree distributions $P(q)$ and different distributions of the shortest-path distances $P(d)$, shown in Figure 2C. We also show the prominent differences in the q-core structure of these networks, cf. Figure 2D.

More importantly, the majority of nodes that belong to the same community in the outbreak phase network appear to be a part of entirely different communities in the immunisation phase network, cf. Figure 3 and the corresponding lists in Supplementary Information. More precisely, we find that only 625 edges established in the outbreak phase persist in the immunisation phase network. They are shown in Supplementary Figure S2 left, in Supplementary Information. A more systematic comparison is made by computing the overlap (Jaccard index defined in Methods) for the correlation networks determined from the successive 2-month intervals, see Supplementary Figure S2, right. The overlap systematically remains below 15%, suggesting that the fluctuation patterns at these intervals can vary between the countries or even provinces within the same country.

2.6 K-Means Clustering and Multi-Fractality of Time Series Within Identified Communities

To further explore the nature of temporal fluctuations of the infection time series of the countries and provinces within each community found using spectral analysis, we apply the K-means algorithm adapted for time series analysis [58], see Methods. It appears that each topological community is further partitioned into several clusters, for example, G1$\cdots$G1$\cdots$4, for the group G1, and so on. Inside each cluster, the corresponding time series have a similar evolution pattern. Hence, the cluster’s typical time series (centroid) is determined for each identified cluster. The results are shown in Figure 5 both for the outbreak and immunisation phase; in the figure legends, the number of countries or provinces belonging to a given cluster is indicated in the brackets in each panel. The names of countries and provinces belonging to each cluster in each group are given in Supplementary Tables S1–S6 in Supplementary Information. Notably, in each network’s group, there is one large and one medium-size cluster. Meanwhile, there are several single-country centroids; as a rule, they indicate a large-population country.

Next we consider the fluctuation function $F_2(n)$ vs. the interval length $n$ for each time series separately, see some examples in Figure 6, and Supplementary Figure S4 in SI. We realised that the similarity of the time series belonging to each cluster manifests itself in the apparent similarity of the slopes of their fluctuation function, which defines the corresponding Hurst exponent. As Figure 6 shows, two different slopes of the fluctuation function can be identified for a majority of time series. At the intervals $n < 14$, a Hurst exponent $0.5 \leq h_2 \leq 1$ can be determined, indicating persistent fluctuations occurring at these time intervals. Meanwhile, an exponent $h_2 > 1$, characteristic to the fractional Brownian motion, is found for $n \geq 14$. In some cases, the determined Hurst exponent reaches values close to two. The histograms of the observed Hurst exponents are shown in Figures 6D,E. 

Compatible with the grouping and different shapes of centroids in the immunisation phase, the distributions of lower and higher values of the Hurst exponents are also different with the increased incidence of the value $h_2 = 0.5$ (white noise), and $h_2 = 2$ (periodic signals) in the immunisation phase. In the following, we show that these large values of the Hurst exponent in many of the studied time series can be related to the occurrence of cyclic trends.

Two prominent examples are shown in Figure 7. The methodology of determining local trends in these time series is described in Methods. The original time series shows a cyclic

### Table 1

|       | $N$  | $E$   | $D$  | $<q>$ | $<C_q>$ | $\#\triangle$ | $\rho$  |
|-------|------|-------|------|-------|---------|----------------|--------|
| Outbreak | 243  | 9,284 | 11   | 3.55  | 38.2    | 0.735          | 58,508 | 0.158 |
| Immunisation | 227  | 5,660 | 7    | 3.10  | 24.9    | 0.666          | 18,857 | 0.112 |
A separate analysis of the fluctuation functions for the trend and the fluctuations around the local trend (detrended signal) reveals that the trend drives the fluctuations beyond the intervals of approximately 14 days; see the insets to Figure 7. The trend has true cyclic fluctuations (the Hurst exponent equals two, within error bars) in the range up to \( n \lesssim 30 \) days. Meanwhile, beyond this range, both the original signal and trend have a lower Hurst exponent in the range \( h_2 \gtrsim 1 \), characterising a fractional Brownian motion. By extending a similar analysis to the above-mentioned typical time series (centroids), we find that they also exhibit cyclic trends but with different cycles characterising different clusters of countries. The corresponding trends are also shown in each panel of Figure 5 as a red line on the top.
of the related centroid. The trend fluctuation functions $F_2(n)$ vs. $n$ shows the cycle characteristics in a large range of the intervals $n$, cf. top left panels of Figure 5. They differ from cluster to cluster and, even for the same country, the cycles also differ in the outbreak and immunisation phase. Generally, larger cycles (in the length and amplitude) are observed in the immunisation phase as compared to the outbreak period, cf. Supplementary Figure S5 in SI. Remarkably, these findings imply that the cycles (or the infection waves) represent an inherent feature of current pandemic which may have some long-lasting consequences.

3 DISCUSSION AND CONCLUSION

In search of universal characteristics of infection dynamics, we have analysed the worldwide empirical data of the SARS-CoV-2 epidemic [44], focusing on the new-infection time series with a daily resolution. The data are purposefully divided into two periods, corresponding to the epidemic’s outbreak and the initial immunisation phase, respectively. Three complementary methods of quantitative analysis have been performed. Specifically, we have analysed the mesoscopic structure of the networks, which embody the significant pairwise correlations among the infection time series of different countries or provinces. The further similarity in the pairs of time series has been analysed by K-means clustering. Finally, the fluctuation function of each time series has been determined using the detrended time series analysis. Our analysis has revealed global clustering and several universal features of the infection dynamics. Our main conclusions are:

- **the worldwide clustering** represented by fourteen temporal patterns of evolution of infection reveals significant similarities transcending geographical regions;
- **the cyclic trends** dominate the infection fluctuations, implying the prevalent infection waves and multi-scale fluctuations around these cycles; typically determined cycles appear in conjunction with the identified clusters;
- **the immunisation phase differs** from the epidemic outbreak phase in all measures considered here, thus quantifying the impact of the (partial) immunisation coverage on the underlying stochastic process and the course of the pandemic.

The mesoscopic (community) structure, as shown in Figure 3 is one of the striking characteristics of the infection-correlation networks; remarkably, it occurs already at zero thresholds, see Supplementary Figure S1 in SI. What comes as a surprise is that these communities constitute almost entirely different nodes (countries or provinces) in the immunisation phase compared to the outbreak phase. Only a few edges established during the outbreak phase persist throughout the entire evolution of the epidemic, as shown in Supplementary Figure S2 in SI. Consequently, the same applies to the contents of the clusters found in these two phases, cf. Supplementary Tables S1–S6 in SI. Notably, a given geographic location and potentially similar cultural and economic development levels, similar healthcare systems and other related factors play some role. However, even such regional groups appear to be a part of a worldwide cluster in both representative phases of the pandemic. Such a picture probably emerges under another dominant driver, common to countries at different locations, and with different cultural and economic developments. In this context, the biology factors, the virus mutations in the interplay with the social behaviour of individuals and groups in the crisis seems to be of the primary importance for the genesis of sustained infection waves, quantified by cyclic trends in different clusters, cf. Figure 7. Our analysis suggests that the waves are ubiquitous in all countries and regions in both representative phases of the pandemic. Meanwhile, the timing, duration and amplitude of these waves vary between different clusters of countries and provinces, likely depending on the applied measures and the corresponding variations in the population behaviours. Moreover, the small-scale fluctuations around these cyclic trends seem to be more region-specific, and depending on the immunisation measures; two comparative examples are shown in Supplementary Figure S5 in SI. A more systematic analysis of these fluctuations and the impact of the immunisation level on the infection dynamics merits future study.

Our analysis of the world infection dynamics of the SARS-CoV-2 pandemic revealed several universal features of the underlying multiscale stochastic processes that go beyond the geographical impact, locally-imposed governmental measures, and partial immunisation phases. Indeed, while these measures are truly valuable for short-term effects, saving lives, and maintaining the functional healthcare system in each country [25], they are much less effective in changing the fundamental nature of the infection process, rooted in the interplay of biology and social behaviours. This work has provided an in-depth analysis of the pandemic’s fundamental phases with an overview that can guide further research into the nature of biosocial interdependencies. The latter factor plays a critical role in the SARS-CoV-2 evolution, where individual biological features of the participants and their role in the collective behaviours need to be better understood. Our effective long-term management of the pandemic and prediction of its future developments rely upon our ability to continue unfolding critical attributes of the underlying biosocial stochastic dynamics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/CSSEGISandData/COVID-19/.

AUTHOR CONTRIBUTIONS

BT, RM, and MM designed research, MM collected data, MM and BT contributed program tools and performed analysis, BT, MM, and RM analysed data, BT produced figures, BT and RM wrote the manuscript, all authors reviewed the manuscript.
**FUNDING**

BT work supported by the Slovenian Research Agency (research code funding number P1-0044). MMD. acknowledge funding provided by the Institute of Physics Belgrade, through the grant by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. RM is grateful to the NSERC and the CRC Program for their support and he is also acknowledging the support of the BERC 2022-2025 program and Spanish Ministry of Science, Innovation, and Universities through the Agencia Estatal de Investigacion (AEI) BCAM Severo Ochoa excellence accreditation SEV-2017-0718, and the Basque Government fund AI in BCAM EXP. 2019/00432.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2022.936618/full#supplementary-material
