Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Dimitriou, P. A., D'Alessandro, M., Chang, B. L., Silvestros, V., Constantinou, E., Pitris, C., Kolios, P., & Van Mieghem, P. (2026). Non-Markovian dynamics and effective reproduction number in COVID-19: Evidence from Cyprus contact tracing data. *PLoS Computational Biology*, *22*(8), e1014578. Article e1014578. https://doi.org/10.1371/journal.pcbi.1014578

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

PLOS Computational Biology

RESEARCH ARTICLE

Non-Markovian dynamics and effective reproduction number in COVID-19: Evidence from Cyprus contact tracing data

Pavlos Alexandros Dimitriou[1,2], Matteo D'Alessandro[3]*, Brian L. Chang[3], Valentinos Silvestros[4], Elisavet Constantinou[4], Costas Pitris[1,2], Panayiotis Kolios[1,5], Piet Van Mieghem[3]

1 University of Cyprus, Department of Electrical and Computer Engineering, Nicosia, Cyprus, **2** KIOS Research and Innovation Center of Excellence, Nicosia, Cyprus, **3** Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands, **4** Ministry of Health, Nicosia, Cyprus, **5** University of Cyprus, Department of Computer Science, Nicosia, Cyprus

* m.d.dalessandro@tudelft.nl

Check for updates

OPEN ACCESS

Citation: Dimitriou PA, D'Alessandro M, Chang BL, Silvestros V, Constantinou E, Pitris C, et al. (2026) Non-Markovian dynamics and effective reproduction number in COVID-19: Evidence from Cyprus contact tracing data. PLoS Comput Biol 22(8): e1014578. https://doi.org/10.1371/journal.pcbi.1014578

Editor: Xiaomin Wan, The Second Xiangya Hospital of Central South University, CHINA

Received: October 9, 2025

Accepted: July 14, 2026

Published: August 21, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014578

Copyright: © 2026 Dimitriou et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution,

Abstract

Using contact tracing data provided by the Cyprus Ministry of Health, infection trees for the first four waves of the COVID-19 epidemic are constructed. In these trees, nodes represent infected individuals, while links indicate the direction of transmission between them. For each infection tree of N nodes, the hopcount distribution from the root node to all other nodes is calculated. The empirical distribution is then compared to the hopcount distribution of infection trees generated by a non-Markovian SI process on a complete graph, with Weibull infection times characterized by a shape parameter α. We compute the values of the shape parameter α that best fit the empirical distribution and find that only values of $\alpha > 1$ are obtained, while the Markovian case is characterized by $\alpha = 1$. A Weibull distributed infection time with shape parameter $\alpha > 1$ is characterized by a unimodal density function with a peak at finite time, consistent with previous findings in the literature. Our analysis therefore suggests that the spreading process is most likely governed by non-Markovian dynamics, and that non-Markovianity can be detected solely from the topology of the infection trees. Finally, we analyze the evolution of the empirical distribution of the number of secondary infections caused by each node in the infection trees across different time windows to estimate the effective reproduction number. In practice, the average number of secondary infections seems to often provide a lower bound of the reproduction number computed by the Cyprus Ministry of Health. When the last level of the trees, composed predominantly of terminal nodes that do not generate further infections, is excluded, the estimate reflects more accurately the dynamics of the epidemic.

and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All data underlying the findings described in the manuscript are available on Zenodo DOI: https://doi.org/10.5281/zenodo.17097764.

Funding: P.V.M., M.D. and B.L.C. have been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant Agreement No. 101019718 (https://cordis.europa.eu/project/id/101019718). P.A.D., C.P. and P.K. have been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 739551 (https://cordis.europa.eu/project/id/739551). P.A.D., C.P., P.K., V.S., E.C. have been funded by the CIPHIS project (Cyprus Innovative Public Health ICT System), C1.1I2, of the NextGenerationEU programme under the Republic of Cyprus Recovery and Resilience Plan (https://commission.europa.eu/projects/cyprus-innovative-public-health-information-and-communications-technology-ict-system-ciphis_en). P.A.D. has been funded by Erasmus+ Project 2023-1-CY01-KA131-HED-000139867 (https://erasmus-plus.ec.europa.eu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

We explore how diseases spread through populations by examining infection trees. Each "node" in the infection tree is a person and each "link" is a transmission that specifies who infected whom during an outbreak. Based on contact tracing data from the COVID-19 epidemic in Cyprus, we develop a method to detect non-Markovianity in disease transmission by analyzing only the structure of these trees. A Markovian process assumes that the chance of spreading a disease depends only on an individual's current state, whereas a non-Markovian process occurs if the timing of infections is influenced by the time passed since the host was first infected. Our results demonstrate that real disease spread exhibits non-Markovian behavior, which can be measured from the structure of infection trees without employing precise chronological information. We also estimate the effective reproduction number, which indicates how many new infections are generated on average by each infected individual. The agreement between our estimates and those reported by the Cyprus Ministry of Health confirms that working with detailed infection tree data is a valid and reliable way to study disease spread. Overall, our work provides a practical framework for understanding disease transmission in real populations.

Introduction

The COVID-19 devastatingly impacted humanity, causing millions of deaths and placing unprecedented strain on healthcare systems worldwide [1]. However, despite these challenges, measurements during the pandemic have provided a unique opportunity to validate epidemiological models and gain deeper insights into disease dynamics. In this work, COVID-19 infection trees, which represent the chains of transmission reconstructed from contact tracing data in Cyprus, are investigated. By analyzing contact tracing information, researchers can uncover transmission patterns of outbreaks, reveal the impact of public health interventions and improve predictive models to better prepare for future outbreaks [2–8]. Cyprus' status as an island, along with the implementation of strict border controls during the epidemic, created a uniquely contained setting that supported comprehensive contact tracing. In a previous work [9], an extensive literature review was conducted comparing studies about COVID-19 infection trees. The study shows that our dataset, spanning multiple epidemic waves, includes more reconstructed transmission trees than any other study to date, highlighting its uniqueness and completeness [9]. This unique scope enables an unprecedented level of analysis, offering valuable insights into the fundamental dynamics of spreading processes in human populations. We emphasize that

assembling such a comprehensive dataset requires substantial resources, which explains why similar large-scale collections are rarely encountered in the literature [6,10,11].

Our main motivation is to measure from the infection trees how much the underlying spreading process departs from the assumptions employed in Markov theory. In Markovian epidemics, all the events in the spreading process happen independently at exponentially distributed random times, and therefore the epidemic itself is memoryless implying that the future states only depend on the actual viral state of the population [12]. Markovian epidemics produce nearly all currently used mean-field epidemic models. Hence, testing whether a Markovian description of reality is valid, is an important achievement. Real-world spreading events are very likely characterized by non-exponential infection and curing times [13–16] and evidence from COVID-19 measures [17,18] suggests that non-Markovian processes [19–22] with unimodal distributed infection times (e.g., Weibull or Gamma distribution) may give a better description of real processes [23].

Numerous studies have calculated different key statistics such as tree size, centrality metrics, the average number of secondary infections and superspreaders from contact tracing data [5,10,24], but little focus on the impact of the non-Markovianity of the spreading process on the tree structure [25]. We employ the equivalence between the non-Markovian susceptible-infected (SI) process in a contact graph and the shortest paths in a graph. In the SI process, susceptible nodes transition to an infected state upon transmission from a neighbor and remain infectious permanently, allowing the pathogen to traverse the full graph's topology. A shortest path tree is equivalent to an infection tree (more details in Sec "Infection trees as shortest path trees") if link weights correspond to infection times, where the infection time is the time needed for a just infected node to infect one of its direct, susceptible neighbors [26]. In our SI process, the time to infect a contact is assumed to be governed by a distribution with shape parameter α, which controls the memory effects in the spreading process. The parameter α is our "non-Markovianity" measure and here we confine to the shape parameter α of the Weibull distribution. When $\alpha = 1$, the infection time follows an exponential distribution, recovering the standard Markovian SI process (S2 Appendix). The Markovian infection trees are uniform recursive trees (URT): at each stage of the process a new node is infected uniformly by one of the existing infected nodes. URTs are characterized by recursive self-similarity [27] because each infection tree can be seen as a union of independent uniform recursive subtrees [28, Sec 16.2.2]. The recursive self-similarity is however "broken" whenever $\alpha \neq 1$, and the tree structure becomes more complicated.

In this work, we extract the parameter α from the empirical hopcount distribution of a set of infection trees of the same number of nodes *N*. The hopcount is the number of links in the shortest path and the hopcount distribution indicates the probability that a path between two randomly chosen nodes in a graph *G* contains a specific number of links, called hops. The hopcount distribution was originally introduced in the early 2000s to study the hopcount of the shortest path between two arbitrary routers in the Internet [29,30]. Indeed, studying the hopcount between arbitrary nodes helps uncovering the topology, optimizing the network infrastructure, and proposing more efficient designs [31]. In network epidemiology [12], where nodes represent infected individuals and links indicate the direction of infection, the hopcount refers to the number of subsequent infections starting from patient zero (root node) to all other nodes in an infection tree.

Methods

Ethics statement

This study was approved by the Cyprus National Bioethics Committee (CNBC 2023.01.146). All data were anonymized, and the requirement for informed consent was waived by the ethics committee. We confirm that all methods were carried out in accordance with relevant guidelines and regulations.

Contact tracing in Cyprus

In Cyprus, the first Health Information System for COVID-19 to support disease surveillance and control, named COVID-19 Emergency Response Platform (COVERP), was developed by the KIOS Center of Excellence at the University of Cyprus, and launched in March 2020 [32]. The timely implementation of this system, resulted in a complete and detailed record of the spread of the disease for the entire country, with 649000 infections in a population of 918100 individuals. Such data, for an entire closed population of almost one million individuals, with well controlled points of access, is not available elsewhere [9]. This situation is unique in Europe and probably internationally as well.

During the first two waves under analysis, contact tracing investigations were performed through telephone interviews from a team of 80 contact tracers. Contact tracers called each confirmed case and conducted investigations to identify the source of the infection and close contacts. For the third and fourth waves, investigations were performed via self-reports, in which confirmed cases declared their own contacts. Close contacts were then informed via SMS and were required to get tested in the following days.

Manual verification and validation were necessary to properly match cases and ensure that transmission trees were correctly reconstructed. The large number of daily tests, as shown in S1 Fig, with more than 80000 tests on some days, demonstrates the extensive testing capacity strengthening confidence that a large fraction of the infected individuals were identified. S2 Fig shows the changes in interventions during these waves providing contextual information on the level of public-health restrictions during the analyzed periods. Obtaining a dataset of this scale and detail is challenging and requires significant financial and human resources. The process of identifying, testing and tracing each case demands extensive coordination between healthcare authorities, laboratories and contact tracing teams. Hence, most countries have not been able to reconstruct enough transmission trees, highlighting the uniqueness of our ensemble.

Dataset

The dataset contains 33358 confirmed infections reported in Cyprus during the first four waves of the epidemic. Each infected individual was assigned a unique case ID, along with information on age, gender, occupation, residence location, date of positive test, and an infector's ID indicating the direction of transmission. However, not all the infected individuals were able to identify their source of infection, resulting in 11577 isolated cases being excluded from the study. Table 1 shows the number of infections with available information about their source of infection, as well as the periods of each wave. The dataset comprises only the 31 days around the peak (maximum number of infected people) of each epidemic wave, and the intervals between waves are not part of the dataset. The time intervals analyzed in this study correspond to periods for which high-quality, epidemiological validated contact-tracing data were available. For the present study, the Ministry of Health provided only the case ID, date of positive test and infector's ID

Table 1. Information about the contact tracing dataset employed in our study for the COVID-19 epidemic in Cyprus. The second column displays the corresponding period. The third column indicates the total number of confirmed cases reported. The fourth column displays the number of cases and the relative percentage of the total cases which belong to an infection tree with at least two nodes.

Wave	Period	Total num. of infected	Num. of Infected in trees of size $N \geq 2$ (%)
Wave 1	Mar 18 – Apr 17, 2020	710	462 (65.1%)
Wave 2	Jul 16 – Aug 15, 2020	314	197 (62.7%)
Wave 3	Dec 13, 2020 – Jan 12, 2021	12345	7527 (61.0%)
Wave 4	Apr 4 – May 4, 2021	19989	13595 (68.0%)

https://doi.org/10.1371/journal.pcbi.1014578.t001

from the dataset. Using this information, infection trees from the first four waves of the epidemic in Cyprus have been constructed.

Infection trees

In the infection trees, the nodes represent infected individuals, while the links indicate the transmission direction between them. Each tree is constructed using data from the 31-day period around the peak (maximum number of daily infected individuals) of each epidemic wave, ensuring that infection chains are not pruned even if some infections occurred outside the studied time frame. Periods of strong epidemic growth are indeed characterized by denser contact-tracing data, and focusing on these intervals facilitates the reconstruction of infection trees. Isolated cases are instances where infected individuals could not provide information on whom they infected or by whom they were infected, and are excluded from the study. The constructed sets of infection trees are sparse and consist of numerous disconnected short transmission chains. Real-world contact-tracing data may contain missing links, due to unreported or untraced contacts. Previous analyses of SARS-CoV-2 transmission networks [7,9,33] show that incomplete information is common and naturally results in multiple disconnected transmission chains rather than a single tree. For the first wave, the whole set of infection trees includes 112 trees, ranging in size from 2 to 31 nodes. The second wave includes 33 trees, with the largest trees consisting of 31 nodes. In the third wave, 2402 trees range in size from 2 to 33 nodes, while the fourth wave involves 4164 trees, with maximum sizes reaching up to 27 nodes [9].

The number and size of reconstructed infection trees vary across epidemic waves. The variation can be attributed to several factors. Waves 1 and 2 were relatively small in overall case numbers, and testing during these waves relied primarily on polymerase chain reaction (PCR) tests, whereas waves 3 and 4 incorporated widespread rapid antigen testing. From Wave 3, the surveillance system also transitioned from telephone-based interviews to electronic documentation, improving the completeness and structure of recorded transmission trees. In addition, transmission patterns may have changed across waves due to the implementation of different control measures, seasonal effects, and the circulation of distinct SARS-CoV-2 variants.

Fig 1 shows the infection network during the fourth wave of the epidemic, in red the infection trees involving more than $N = 5$ infected individuals and in gray trees representing those with $N < 5$. The zoomed-in views highlight an example of the transmission chains with $N = 14$, where the root node (or patient zero) is at the top of the tree, and subsequent nodes are organized into levels representing the depth of the infection tree.

The hopcount defines the number of hops (links) along the unique path from the root to a specific node in the infection tree. By calculating the hopcount for every node, we can determine the level set, which represents the group of nodes across discrete distances from the root. Fig 2 displays an example of the hopcount of all the nodes belonging to a tree of size $N = 14$ in the dataset.

Data limitations

The infection trees in the dataset are affected by intrinsic limitations:

1. The available contact tracing data only provide partial subgraphs of the global infection tree, offering an incomplete view of the overall transmission dynamics.
2. Contacts of infected individuals who tested negative are not reported, and only contacts who subsequently tested positive are included.
3. Given the limited time window selected, each infection tree only includes individuals which have been infected once.
4. The dataset does not distinguish between imported cases and local cases with unidentified sources; consequently, any individual without a traceable parent is treated as a new root node.

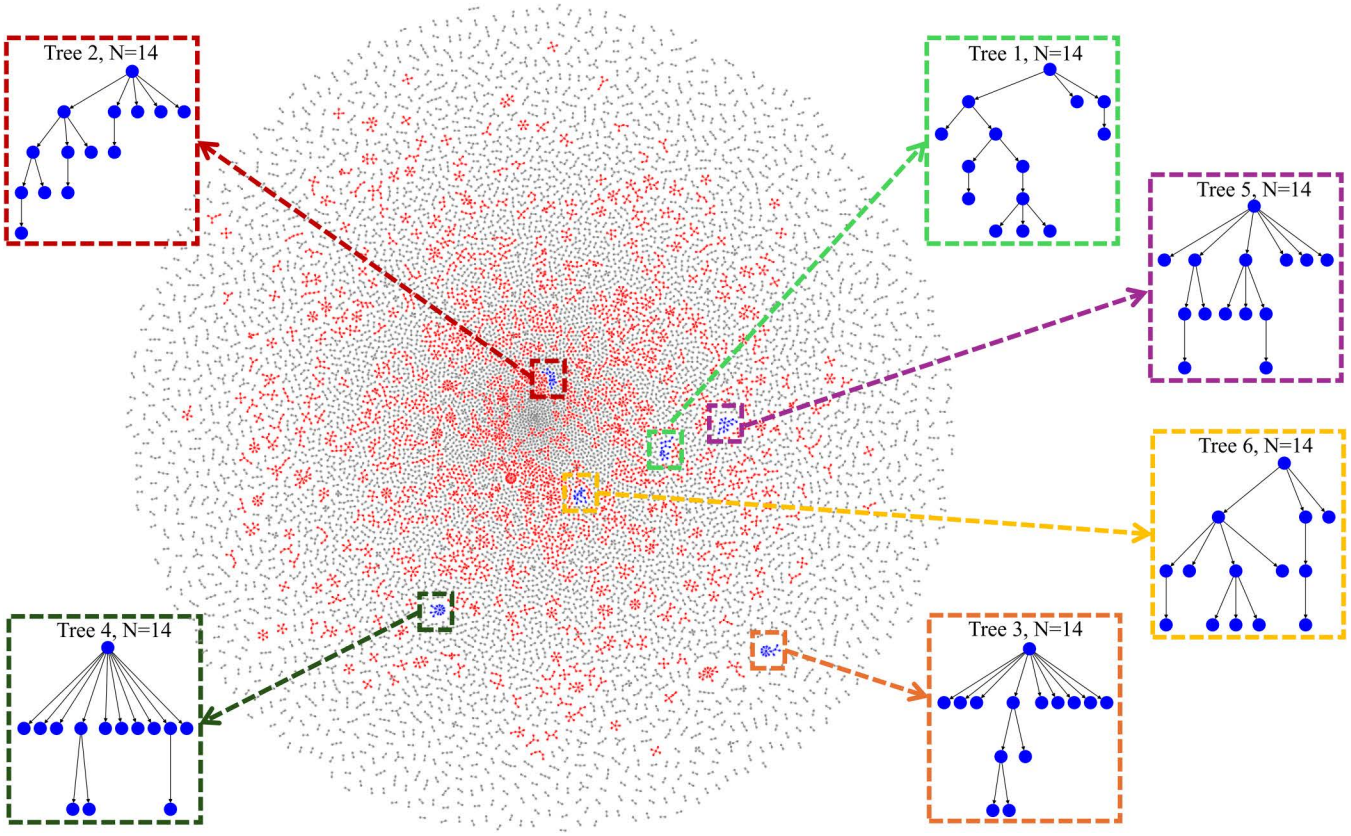


Fig 1. Infection trees of the fourth wave (April 2021 - May 2021) of the COVID-19 epidemic in Cyprus. The zoomed-in view displays a subset of the infection trees of size $N = 14$. Trees with less than 5 nodes are shown in gray, while larger trees (5 or more nodes) are shown in red.

https://doi.org/10.1371/journal.pcbi.1014578.g001

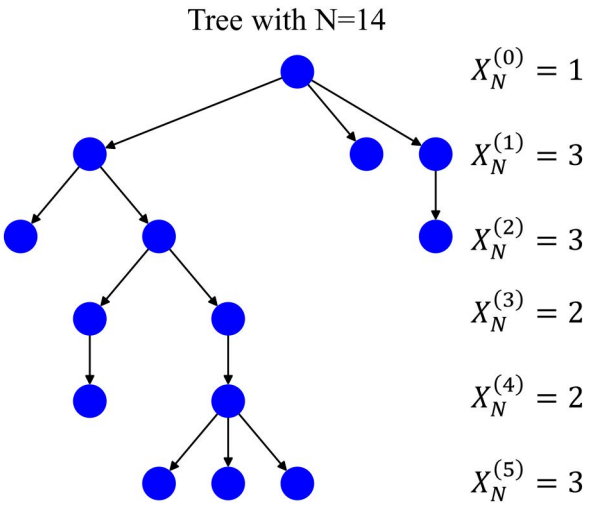


Fig 2. A real infection tree of the COVID-19 epidemic in Cyprus with $N = 14$ nodes. The variables $X_N^{(h)}$ represent the number of nodes located exactly at h hops from the root. For instance, at hopcount $h = 1$, there are $X_{14}^{(1)} = 3$ nodes.

https://doi.org/10.1371/journal.pcbi.1014578.g002

Theory of infection trees

Infection trees as shortest path trees. When analyzing an epidemic spreading on a graph G without re-infections, an infection tree (also known as transmission tree) is a directed acyclic subgraph of G whose nodes are the infected individuals and the links show "who infected whom," starting from an initial case, the root or "patient zero" [34]. Two main properties of infection trees are investigated: (I) the distribution of the hopcount (i.e., the number of links along the shortest path) between the root and any other node in the infection tree; (II) the distribution of the outdegree of the nodes in the infection tree.

We employ the analogy between the shortest path problem on a graph G with independent and identically distributed (i.i.d) link weights and a susceptible-infected (SI) spreading process on G (see S2 Appendix for a rigorous definition of the Markovian process), where the infection times possess the same distribution as the link weights in the graph. The equivalence follows by viewing the transmission from the root to any node as finding the shortest path between them: assigning an independent infection time to each contact, the time at which an individual becomes infected is the minimum total infection time over all paths from the root, i.e. a shortest-path distance. Under the SI assumptions (no recovery/reinfection), the resulting parent-child relations form the corresponding shortest-path tree. The shortest path tree (SPT) [28, Chapter 16] is the union of the shortest paths from a source node to a set of other nodes in the graph G.

Markovian SI infection tree and shortest path tree with exponential link weights. The shortest path tree in the complete graph K_N with i.i.d exponentially distributed link weights is exactly equivalent to the infection tree of a homogeneous Markovian SI spreading process on K_N, where the link weights correspond to the exponentially distributed infection times (see S2 Appendix) with mean infection time $1/\beta$. In our definition, an infection time is the random duration drawn for a transmission attempt along a specific link. Because the epidemic spreads through competing processes, a susceptible individual is ultimately infected by the neighbor whose drawn infection time is the shortest. Indeed, let us suppose that the root node A is infected at time $t=0$ and that the infection rate of the process is $\beta = 1$ for all the links between infected and susceptible nodes. The next node in the infection tree, which we call B, is the one infected first by A whose infection time is the smallest. The infection times for all the other nodes, which are infected later in the process either by the root or by another infected node, are still exponentially distributed with rate β due to the memoryless property of the exponential distribution [28, p. 400–401]. Indeed, if node A infects node B at time u, the probability that node A infects another node C after time $t+u$ is in general [28, p. 112]

$$\begin{aligned}\Pr[T_{A\to C} > t+u | T_{A\to C} > u] &= \frac{\Pr[\{T_{A\to C} > t+u\} \cap \{T_{A\to C} > u\}]}{\Pr[T_{A\to C} > u]} \\ &= \frac{\Pr[T_{A\to C} > t+u]}{\Pr[T_{A\to C} > u]}\end{aligned} \tag{1}$$

where $T_{A\to C}$ is the random time at which A tries to infect C. If the infection times are exponentially distributed then (1) becomes

$$\Pr[T_{A\to C} > t+u | T_{A\to C} > u] = \frac{e^{-\beta(t+u)}}{e^{-\beta u}} = \Pr[T_{A\to C} > t]$$

and the probability that node A infects another node C after time $t+u$ is equal to the probability that node A infects another node C after time t. Additionally, the minimum between independent exponential random variables is still an exponential random variable with rate equal to the sum of the single variable rates [28, p. 54]. Thus, the next infection in the Markov process will always happen at an exponentially distributed random time. Given that in a Markovian SI process all the infection link processes are independent, the sum of the infection times in the path of the SI infection tree between the root A and a given node D, is exactly the weight of the shortest path between A and D where the link

weights are i.i.d exponentially distributed random variables with rate β. We can thus employ the well-known theory of the shortest path trees in the complete graph [28, Sec. 16.2], to devise the properties of the infection trees of a Markovian SI process on K_N. In particular, we exploit the fact that the shortest path tree in the complete graph with exponential link weights is a uniform recursive tree [28, Sec. 16.2]. A uniform recursive tree (URT) of size N is a random tree rooted at a given node A. At each stage a new node is attached uniformly to one of the existing nodes until the total number of nodes is equal to N. Due to the memoryless property of the exponential distribution and with all the infection link rates equal to β, the Markovian SI infection tree on the complete graph is a uniform recursive tree.

The hopcount h_N in the URT of N nodes is defined as the smallest number of links between the randomly selected root A and a destination chosen uniformly from all the other nodes $\{1, ..., N-1\}$. The exact probability density function of the hopcount in the URT with N nodes is [28, Corollary 16.3.2]

$$\Pr[h_N = k] = \frac{(-1)^{N-(k+1)} S_N^{(k+1)}}{N!}, \tag{2}$$

where $S_N^{(k)}$ indicates the Stirling numbers of the first kind [35, 24.1.3]. The distribution (2) describes the hopcount distribution in the infection trees of a Markovian SI process on a complete graph with N nodes. For large N, to the first order, (2) is well approximated by a Poisson distribution [28, Eq. 16.14]

$$\Pr[h_N = k] \sim \frac{(\log N)^k}{Nk!}. \tag{3}$$

Non-Markovian SI infection tree and shortest path tree with non-exponential link weights. Let us consider the shortest path tree in a weighted complete graph K_N with N nodes. If each link has an assigned random weight described by the random variable $W \overset{d}{=} X^{1/\alpha}$, where X is an exponential random variable $\Pr[X \leq x] = 1 - e^{-x}$ with mean 1 and where $\alpha > 0$ is a positive shape parameter, then the distribution of the link weights corresponds to

$$\Pr[X^{1/\alpha} \leq x] = \Pr[W \leq x] = 1 - e^{-x^\alpha}. \tag{4}$$

The Weibull distribution is [28, Sec. 3.5.3]

$$F_W(x) = \Pr[W \leq x] = 1 - e^{-(\beta x)^\alpha}. \tag{5}$$

and the related probability density function reads

$$f_W(x) = \alpha\beta(\beta x)^{\alpha-1} e^{-(\beta x)^\alpha}. \tag{6}$$

Equation (4) implies that the single link weights are distributed with a Weibull distribution with shape parameter $\alpha > 0$ and rate parameter $\beta = 1$. Fig 3 shows the different properties of the Weibull density function between α smaller and bigger than 1. For $\alpha < 1$, the density function diverges at zero and becomes more heavy-tailed. For $\alpha > 1$, we observe a function with a single peak at $x > 0$ which has a root in the origin.

The hopcount distribution $\Pr[h_N = k]$ (for integers $k = 0, ..., N-1$) of the shortest path tree with i.i.d. Weibull link weights is not known explicitly for any finite N, but has been computed [29, Sec. 6.1] asymptotically [29,36] for $N \to \infty$ as:

$$\Pr[h_N = k; \alpha] \approx \frac{1}{N} \frac{(\log \alpha N)^{\alpha k}}{\Gamma(\alpha k + 1)}, \tag{7}$$

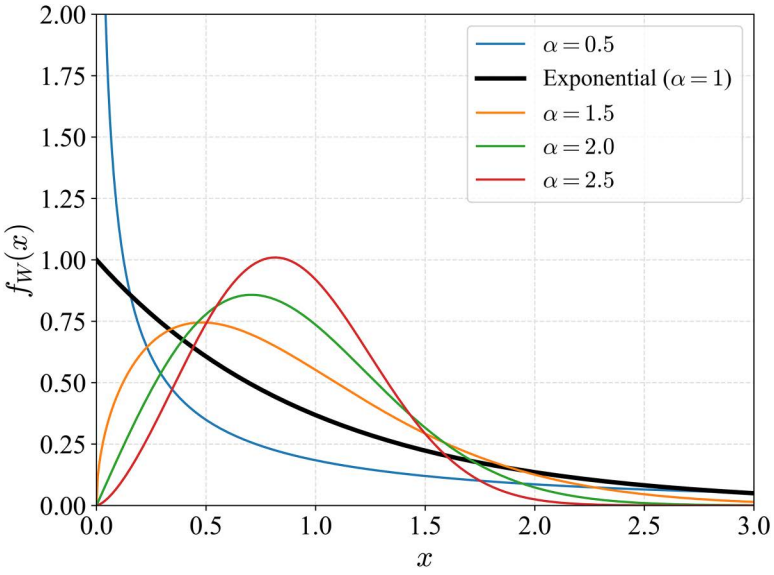


Fig 3. The Weibull probability density function (6) with rate parameter β equal to 1 and different shape parameters $0.5 \leq \alpha \leq 2.5$.

https://doi.org/10.1371/journal.pcbi.1014578.g003

which depends on the tree size N and on the shape parameter α. For $\alpha = 1$ equation (7) reduces to the Poisson distribution in (3). While in the Markovian case the relation between the shortest path trees and the infection trees is a natural consequence of the memoryless property of the exponential distribution and of the minimum between independent exponential random variables (see Sec "Markovian SI infection tree and shortest path tree with exponential link weights"), in a non-Markovian SI process on a graph the exact computation of the statistical properties of the infection trees is complex.

We replace the exponential distribution in the Markovian SI process presented in Sec "Markovian SI infection tree and shortest path tree with exponential link weights" with a Weibull distribution which is usually employed in non-Markovian models for spreading on networks [19–22]. The shape parameter α gives an indication of the non-Markovianity of the process [37] as for $\alpha = 1$ the pdf (6) is just an exponential with rate β. Contrary to the Markovian case $\alpha = 1$, the SI infection tree on the complete graph with Weibull infection times is not a uniform recursive tree. Suppose node A is infected at time 0 and it tries to infect its neighbors B and C. If the infection attempt towards B is successful the random infection time $T^{(\alpha)}_{A\to B} = u$ and $T^{(\alpha)}_{A\to C} > u$ because the infection of node C from A will certainly happen after a time u. We thus define a new random infection time from A to C, conditioned on the event $T^{(\alpha)}_{A\to C} > u$, as the residual random variable $R^{(\alpha)}_{u;A\to C} = T^{(\alpha)}_{A\to C} - u$, whose distribution function is

$$\Pr[R^{(\alpha)}_{u;A\to C} \geq t] = \Pr[T^{(\alpha)}_{A\to C} > t + u | T^{(\alpha)}_{A\to C} > u] = \frac{e^{-(\beta(t+u))^{\alpha}}}{e^{-(\beta u)^{\alpha}}}, \tag{8}$$

having employed (1) with the Weibull distribution (4). For $\alpha \neq 1$, the infection probability depends on the history of the infections and the process is clearly non-Markovian.

The process of building the shortest path tree in a complete graph K_N with Weibull distributed i.i.d weights, produces a shortest path tree without the symmetry of the uniform recursive tree in the exponential weights setting. Indeed, if the link weight $w_{A,B}$ from the root A to a given node B is the smallest among all the links from A, we are sure that the link weight from A to another node C is such that $w_{A,B} < w_{A,C}$. Choosing $w_{A,B} = u > 0$ and knowing that $w_{A,C}$ is Weibull distributed, implies that

$$\Pr[w_{AC} > t + u | w_{AC} > u] = \frac{e^{-(\beta(t+u))^{\alpha}}}{e^{-(\beta u)^{\alpha}}} \neq \Pr[w_{AC} > t] = e^{-(\beta t)^{\alpha}},$$

equivalent to the expression (8) obtained for the non-Markovian SI process. With non-exponential link weights, the process of building the shortest path tree on a graph is equivalent to the non-Markovian SI process on the same topology.

Reproduction number and infection trees

The basic reproduction number R_0 is defined [38] as the expected number of secondary cases produced, in a completely susceptible population, by a *typical* infected individual during its entire period of infectiousness. After sufficiently long time, a non-zero fraction remains infected if $R_0 > 1$, else the epidemic will die out if $R_0 < 1$ [12,39]. A rigorous computation of the R_0 requires the definition of a *next generation operator*, which describes how many secondary cases arise from an individual which generally belongs to a heterogeneous population characterized by a given distribution. The basic reproduction number R_0 is defined as the dominant eigenvalue of the next generation operator [38]. Alternatively, R_0 can be defined from compartmental disease transmission model based on a system of differential equations [40].

The classic definition of the basic reproduction number usually ignores the underlying contact network between individuals in a population [28, 40, Sec 17.1] and can often be replaced by the epidemic threshold. The epidemic threshold τ_c represents the strength of the epidemic process and depends on the underlying contact network. In the mean-field approximation $R_0 = \tau\lambda_1$, where λ_1 is the spectral radius of the adjacency matrix of the underlying contact network [28, Sec. 17.3].

Estimating the basic reproduction number R_0 or the epidemic threshold τ_c from real-world data is complex, because it requires a detailed understanding of various population characteristics, such as age, sex, and other demographic factors, as well as their influence on disease transmission. Additionally, it necessitates the knowledge of the full contact network between individuals, the temporal evolution of the number of susceptible individuals, and potential external influences like vaccination rates or behavioral changes in response to the epidemic [41]. Indeed, when the basic reproduction number R_0 is computed from compartmental models, many assumptions [38,40,42] are needed.

If infection trees are available, then the reproduction numbers can be estimated from data itself [4–6,24] with minimal assumptions. In an active epidemic, it is often more practical to employ the effective reproduction number $\hat{R}$, defined as the actual average number of secondary cases generated per primary case [5]. For any individual i within an observed infection tree, the number of secondary infections is directly represented by their outdegree. By considering a specific temporal window $W_{\Delta t} = (t - \Delta t, t]$, we can compute the outdegree distribution for all cases tested positive within $W_{\Delta t}$. The mean of this distribution provides the empirical effective reproduction number $\hat{R}(t; \Delta t)$ for that period. The knowledge of the full outdegree distribution also allows to extract information about the variance of the individuals' infectivity of the spreading process, often overlooked when considering average quantities [24].

Modeling assumptions

In view of the limitations (Sec "Data limitations") that characterize the infection trees in the dataset, we propose the following assumptions to model the data using our theory (Sec "Theory of infection trees"):

1. Each infection tree of size N nodes is an independent realization of the non-Markovian SI process on a complete graph K_N with N nodes (Sec "Non-Markovian SI infection tree and shortest path tree with non-exponential link weights"). The infection time from an infected to a susceptible node is assumed to be Weibull distributed with

probability density function (6), rate parameter $\beta = 1$ for each link and shape parameter α. In our definition of the non-Markovian SI epidemic, each infection process is assumed to be independent from the other infection processes in the network.

2. Each infection tree is assumed to represent a complete spreading process. If the tree has size N, we assume that the infection tree is the resulting transmission chain within a fully connected community of N individuals, where one individual is initially infected and the spreading process continues until all N individuals have become infected.

Overall, our framework provides a minimal model to detect and characterize non-Markovian dynamics in the empirical spreading process.

Estimate of non-Markovian dynamics from infection trees

In this section we summarize the two methods employed to measure the non-Markovianity of the infection trees through the estimation of the parameter α from the infection tree structure. A more comprehensive description is given in S4 Appendix. All the simulated infection trees are generated using the Dijkstra's algorithm [43], as explained in S3 Appendix.

Method 1. In each wave we create a set S aggregating trees of the same size N and for each set S we compute the hopcount distribution

$$\hat{\Pr}[h_N = k|\text{data}] = \frac{1}{N}\frac{1}{n_{\text{trees}}}\sum_{i=1}^{n_{\text{trees}}}(X_N^{(k)})_i, \tag{9}$$

where $(X_N^{(k)})_i$ is the level set at hopcount k of the i-th tree of size N in the set S. We then fit the empirical hopcount distribution with (7) extracting the non-Markovianity parameter α with a least square method. Each set of trees is resampled with replacement $N_{\text{boot}} = 1000$ times obtaining N_{boot} new resampled set of infection trees of the same size N. For each resampled set S_j, with $j = 1, ..., N_{\text{boot}}$, a least squares estimate $\hat{\alpha}^{(j)}_{\text{data}}$ of the shape parameter is obtained. From the distribution of the N_{boot} estimates of the non-Markovianity parameter α, we extract the average, the 5th and the 95th percentiles. The three values are then adjusted comparing the least square estimates computed with (7) from the data, with the least square estimates of α computed on simulated set of infection trees. The simulated trees are obtained from the simulation of the SI process on a complete graph with fixed N and selected shape parameter $\alpha_{\text{sim}} \in [0.5, 4.5]$ for the infection time. Fig 4 presents an overview of the method.

Method 2. In each wave we aggregate trees of the same size N and we compute the empirical outdegree distribution of the roots of the trees. We resample the trees of same size N with replacement and we derive $N_{\text{boot}} = 1000$ bootstrapped root outdegree distributions. For each resampled distribution ($j = 1, ..., N_{\text{boot}}$), we make a maximum likelihood (see Eq. (S9) in S4 Appendix) estimate $\hat{\alpha}^{(j)}_{d_r}$ of the shape parameter α, maximizing the log-likelihood of observing the resampled empirical root outdegree distribution given the simulated distribution $\hat{\Pr}[d_{\text{root}} = k|\alpha_{\text{sim}}]$ (with α_{sim} values in the range [0.5,4.5]). The average and the percentiles of the N_{boot} estimates give the final value $\hat{\alpha}_{d_r}$ of the shape parameter α with confidence intervals. Fig 5 presents an overview of the method.

To ensure the reliability of our methods, we performed a comparative study of their statistical power and robustness (see S4 Appendix and S5 Fig). The analysis confirms that both methods accurately track the underlying shape parameter α and exhibit comparable statistical power (probability of rejecting the Markovian hypothesis given a true non-Markovian set of data).

Effective reproduction number from the evolution in time of the outdegree distribution

We examine the distribution of outdegrees in infection trees, that is the number of secondary cases attributed to each infected individual. Analyzing the distribution within a given time window captures how the infection trees reflect the evolution

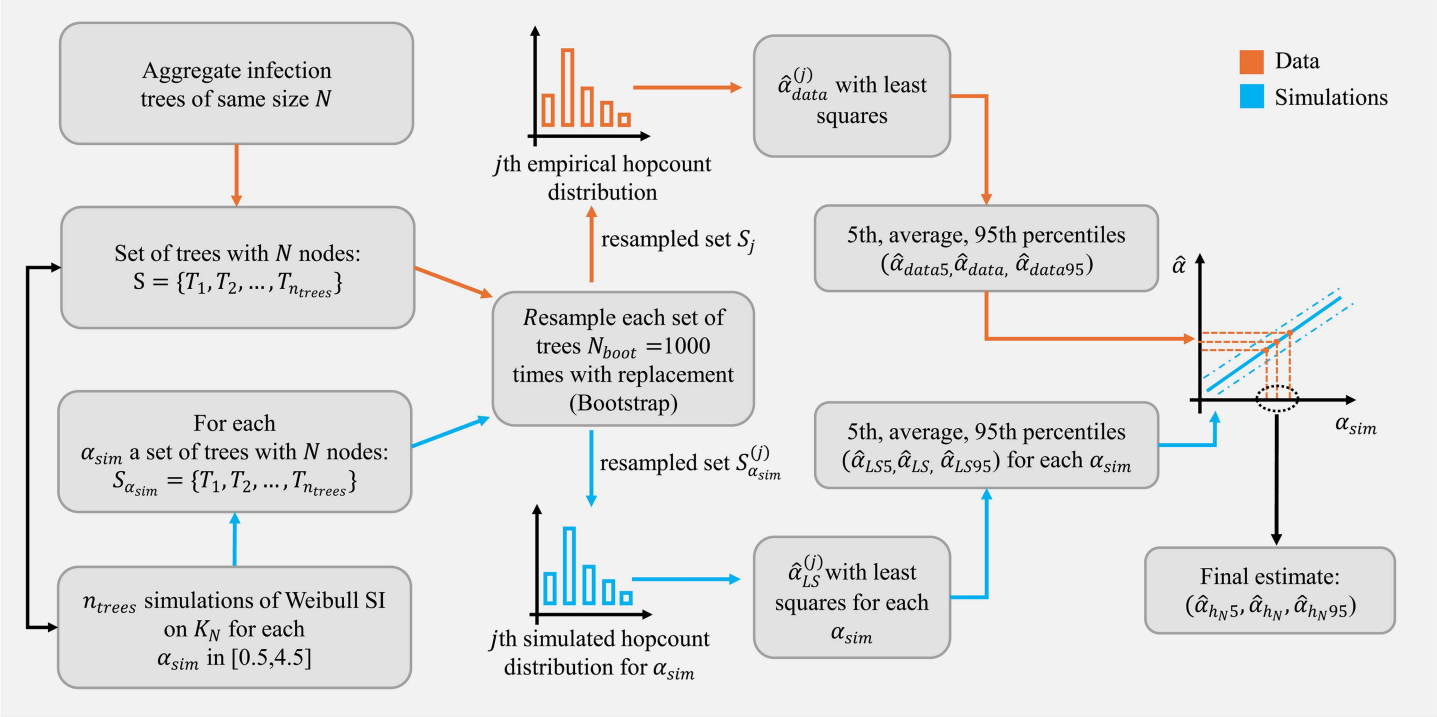


Fig 4. Flow chart of Method 1 based on hopcount distribution, details in S4 Appendix.

https://doi.org/10.1371/journal.pcbi.1014578.g004

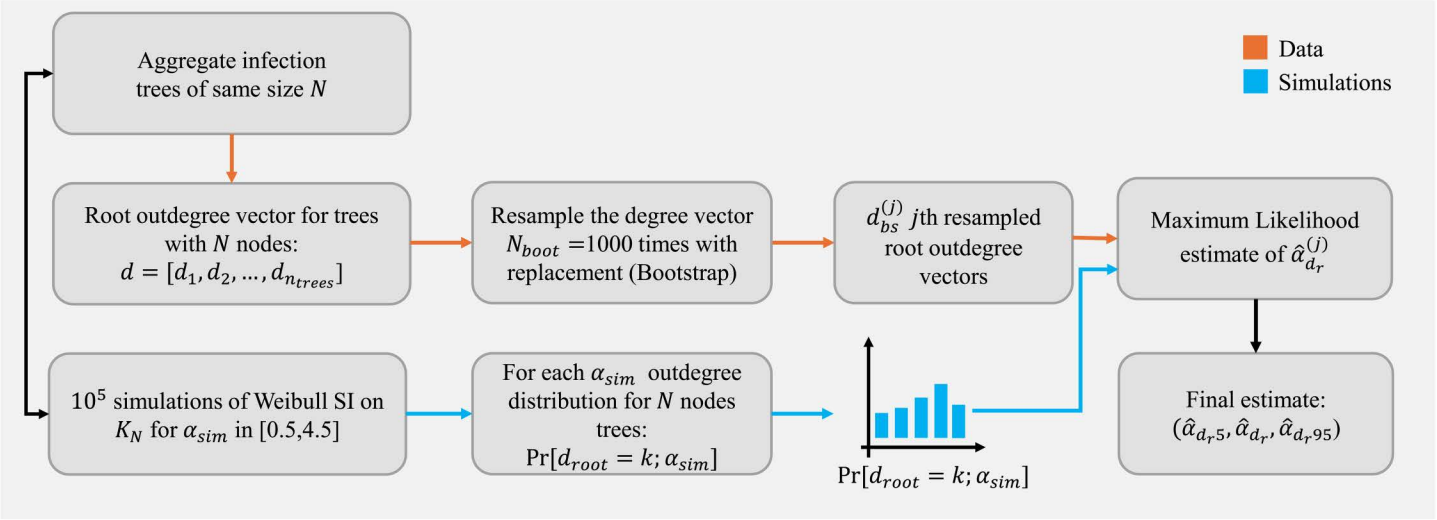


Fig 5. Flow chart of Method 2 based on root outdegree distribution, details in S4 Appendix.

https://doi.org/10.1371/journal.pcbi.1014578.g005

of the epidemic spreading. We define the (dynamic) effective reproduction number $\hat{R}(t; \Delta t)$ as the average number of secondary infections caused by individuals who were tested positive within the time window $W_{\Delta t} = (t - \Delta t, t]$. Such quantity enables direct comparison with reproduction number estimates obtained from alternative modeling approaches applied to the

same epidemic in Cyprus [42,44]. We complement the effective reproduction number $\hat{R}(t;\Delta t)$ with a 95% confidence interval obtained resampling $N_{\text{boot}} = 1000$ times the outdegree distributions of the confirmed cases in the time window $W_{\Delta t}$.

Our definition of the dynamic effective reproduction number $\hat{R}(t;\Delta t)$ is not suitable for real-time monitoring purposes. This is because its computation relies on complete information about the secondary cases generated by individuals who tested positive within the selected time window, including infections that may have been recorded after the window has closed. A real instantaneous reproduction number would require taking into account the period of infectivity of a single individual and its contact network (or at least its average number of contacts), which would necessitate additional assumptions and information that are not available for our dataset.

Because our dataset contains no periods of unmitigated transmission without external intervention (S2 Fig), we cannot isolate the basic reproduction number R_0 under natural conditions. Consequently, this study focuses exclusively on the effective reproduction number $\hat{R}$, which reflects actual transmission dynamics.

Results

Infection trees show non-Markovian properties

Fig 6 shows the hopcount distribution of the infection trees in wave 4 with size $N = 5,6,7,8$. These tree sizes represent the most frequently observed sizes in the empirical data. In red the empirical hopcount distribution is displayed. In dashed blue the distribution with parameter $\hat{\alpha}_{\text{data}}$, that is the average of the estimates of $\hat{\alpha}^{(j)}_{\text{data}}$ with $j = 1,\ldots,1000$ obtained on the resampled empirical distributions. The blue area fills the space between the two distributions with parameters equivalent to the 5th and the 95th percentiles of the distribution of the bootstrapped estimates.

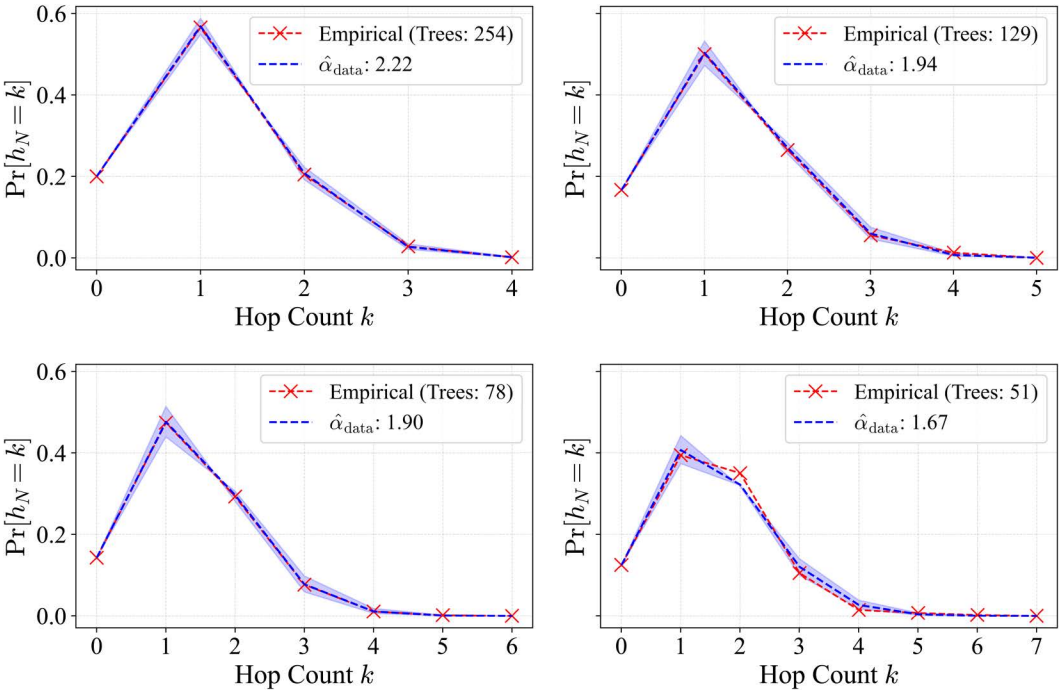


Fig 6. Hopcount distribution for infection trees belonging to wave 4 of size $N = 5,6,7,8$. Red line: empirical hopcount distribution obtained aggregating trees of the same size N. Dashed blue line: distribution (7) with parameter $\hat{\alpha}_{\text{data}}$, which corresponds to the average of the least squares estimates $\hat{\alpha}^{(j)}_{\text{data}}$, $j = 1,\ldots,1000$. The blue area spans the distributions (7) with shape parameter α in the 90% confidence intervals of the least square estimate $\hat{\alpha}_{\text{data}}$.

https://doi.org/10.1371/journal.pcbi.1014578.g006

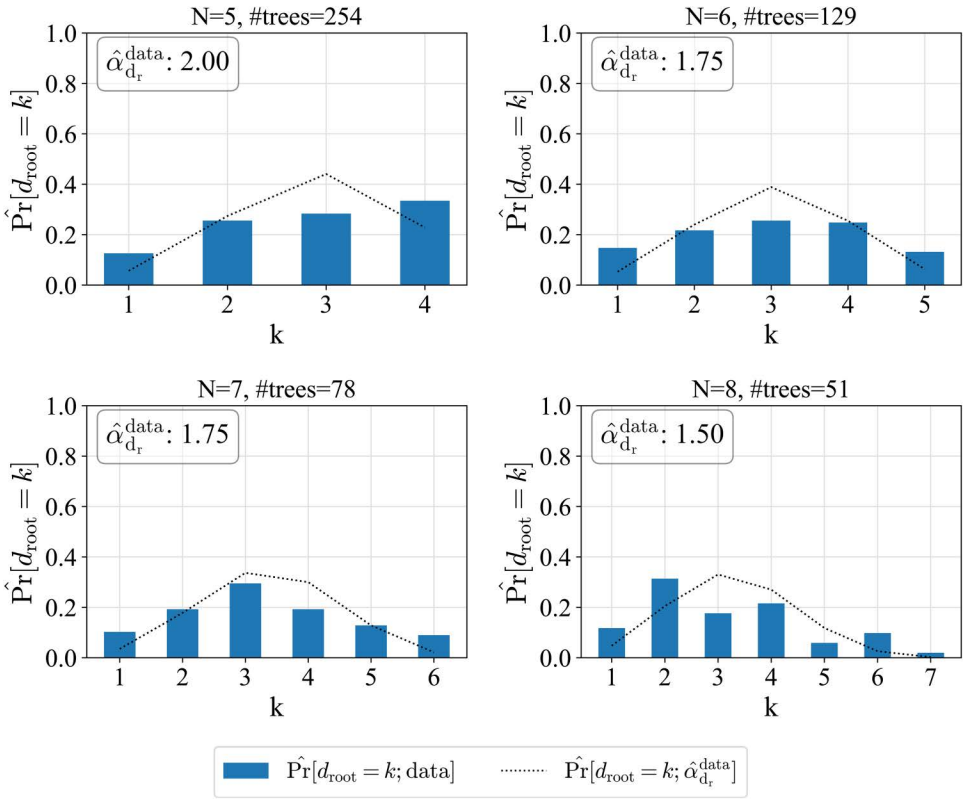


Fig 7. Empirical root outdegree distribution obtained from the set of infection trees of size *N* = 5,6,7,8 belonging to wave 4. In black the root outdegree distribution obtained from the non-Markovian SI simulation on the complete graph with size *N* and α_{sim} equal to the maximum likelihood estimate $\hat{\alpha}_{d_r}^{data}$ in the range [0.5,4.5] (with steps of 0.25) given the empirical distribution. In blue the empirical root outdegree distribution.

https://doi.org/10.1371/journal.pcbi.1014578.g007

The least squares estimates $\hat{\alpha}_{data}$ extracted from the empirical distribution are then corrected employing the hopcount distribution resulting from the simulation of the SI process. The detailed correction procedure is displayed in S3 Fig.

The second method to measure the non-Markovian parameter α (Fig 5) does not use the approximation in (7), but only relies on the root outdegree distribution from the SI process simulations (S4 Fig). Fig 7 shows the empirical distribution for wave 4 and tree sizes $N = 5,6,7,8$. The black dotted line is the root outdegree distribution obtained from simulations with α_{sim} equal to the maximum likelihood estimate $\hat{\alpha}_{d_r}^{data}$ of the shape parameter, given the root outdegree distribution of data (in blue).

In Fig 8, we show the estimate of the non-Markovianity parameter α for all the available trees with $N > 4$, $n_{trees} > 3$ and for all the waves. The threshold of at least three independent chains per tree size was imposed to ensure meaningful statistical comparison (see also S5 Fig). Wave 2 is excluded from Fig 8 because it represented a relatively small epidemic wave and did not contain a sufficient number of infection trees of each size to meet this minimum sample requirement.

The error bars show the limits of the 90% confidence intervals devised with the bootstrap resampling. Fig 8 demonstrates that the average values of the estimate of α are always greater than 1 regardless of the tree size and of the number of trees in the dataset. The confidence intervals are clearly dependent on the size of the single tree set and increase when we have a small number of trees in a set. In the cases in which we have a larger number of trees, both the methods give a value of α greater than 1 with 95% statistical significance. Due to the limited size of the specific tree sets, the points for which the Markovian hypothesis cannot be rejected (for both methods) with 95% statistical significance are: wave 1

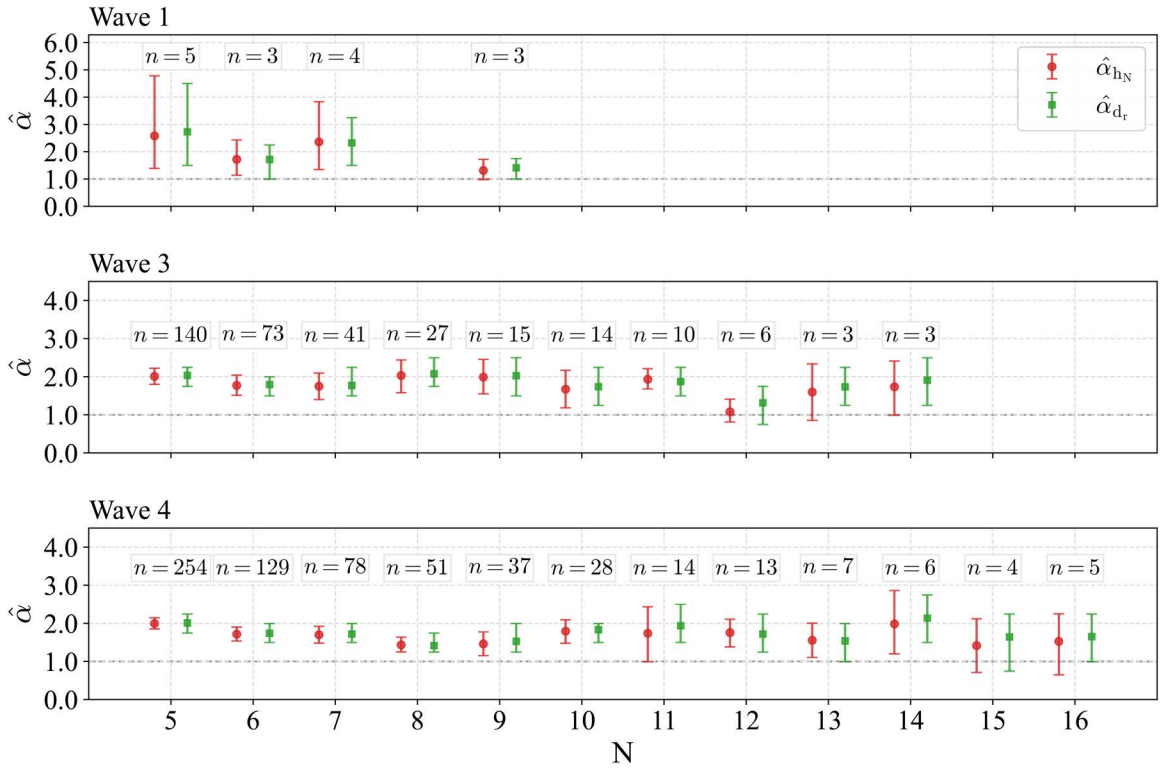


Fig 8. Summary of all the parameters α obtained on the sets of infection trees of wave 1, 3 and 4. In red the estimates of the shape parameter α employing the method based on the hopcount distribution of the infection trees ($\hat{\alpha}_{h_N}$). In green, the estimates of the shape parameter α employing the method based on the root outdegree distribution of the infection trees ($\hat{\alpha}_{d_r}$). The error bars represent the 90% confidence intervals given by the bootstrap procedure performed resampling 1000 times with replacement each set of trees of the same size. The number *n* on top of each estimate indicates how many infection trees are employed for a given size *N*.

https://doi.org/10.1371/journal.pcbi.1014578.g008

tree size $N=9$; wave 3 tree sizes $N=12$; wave 4 tree sizes $N=15, 16$. However, the average value of α is bigger than one also in these cases. Finally, the values of our non-Markovianity parameter α seem to always be around 1.5-2, regardless of the tree size or wave, meaning that the non-Markovian model we have employed uncovers some *universal* properties of the measured spreading process. The values of α bigger than 1 are also in line with previous evidence from COVID-19 literature [17,18], which suggest that infection times can be fitted (see also S6 Fig) with a Weibull distribution with shape parameter bigger than one [37].

To verify whether the variability in the confidence intervals width stems from the methods itself or from the data, we performed a comparative analysis of both methods on synthetic data (S5 Fig). Our synthetic analysis demonstrates that both methods are remarkably robust and with comparable statistical power. The confidence intervals observed in Fig 8 are wider than the one obtained on synthetic data (S5 Fig), because they reflect the natural noise and variance present in real-world contact tracing data. Hence, the "empirical" confidence intervals provide an upper bound which allows us to safely reject the Markovian hypothesis whenever the 5th percentiles of the estimates in Fig 8 are above $\alpha = 1$.

Effective reproduction number

From the outdegree distribution of the infection trees in different time windows, we can make an estimate of the expected number of secondary cases produced by an individual tested positive in that specific time period. We employ the definition of the effective reproduction number $\hat{R}(t;\Delta t)$ within the time window $(t-\Delta t, t]$ presented in Sec "Effective reproduction

number from the evolution in time of the outdegree distribution" to examine the evolution of the COVID-19 epidemic in Cyprus. We then compare our effective reproduction number $\hat{R}(t;\Delta t)$ with previous measures of the reproduction number for the same epidemic in Cyprus [42,44].

Figs 9 and 10 display the number of confirmed cases of SARS-CoV-2 transmission during the first two epidemic waves in Cyprus, while S9 and S10 Figs during the third and fourth waves. The upper panels show violin plots of the daily outdegree distribution for the cases tested positive on that single day. The width of each violin reflects the variability in transmission, while the color intensity corresponds to the mean outdegree $\hat{R}(t;1\text{ day})$ for that day. Lighter colors indicate lower average secondary transmissions, and darker shades indicate more intense transmission clusters. The central panels display the corresponding number of daily reported cases during the same period. Peaks in the number of daily cases often correspond to broader outdegree distributions with higher average, suggesting super-spreading events or days with higher transmission heterogeneity. In the bottom panel of Fig 9, we compare our effective reproduction number $\hat{R}(t;1\text{ day})$ for the first wave with the reproduction number $\hat{R}_t$ reported in [42]. The estimate in [42] was derived using a meta-population model proposed by Peng et al. [45], a generalization of the classical SEIR (susceptible-exposed-infected-recovered) model. For each day, we apply a bootstrap resampling method to generate a distribution of outdegree values, from which we calculate the mean and the 95% confidence intervals. The red points in the figure represent the bootstrapped means, with the red area indicating the 95% confidence intervals. The blue line shows the effective

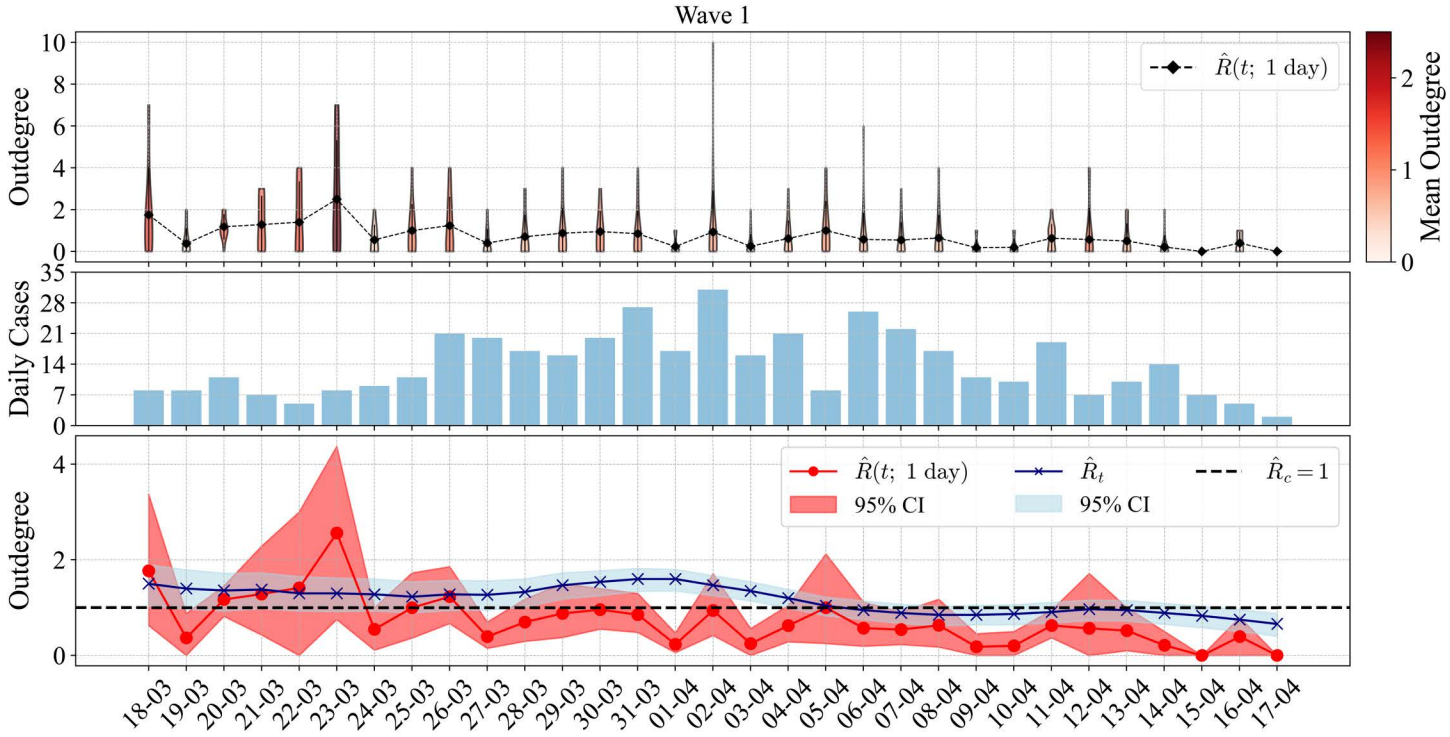


Fig 9. Daily evolution of the first wave of the COVID-19 epidemic in Cyprus. Daily out-degree distribution (top) and daily reported case counts (middle) during the first COVID-19 wave in Cyprus. Violin plots represent the distribution of secondary infections per case per day; color intensity indicates the mean outdegree. (Bottom) Evolution of the effective reproduction number $\hat{R}(t;1\text{ day})$ in the first wave of the COVID-19 epidemic in Cyprus. The red line shows the mean outdegree for each day derived from bootstrapped samples, with the 95% confidence intervals indicated by the red area. The blue line represents the effective reproduction number reported in [44]. In dashed black the reference line at the critical value $\hat{R}_c = 1$.

https://doi.org/10.1371/journal.pcbi.1014578.g009

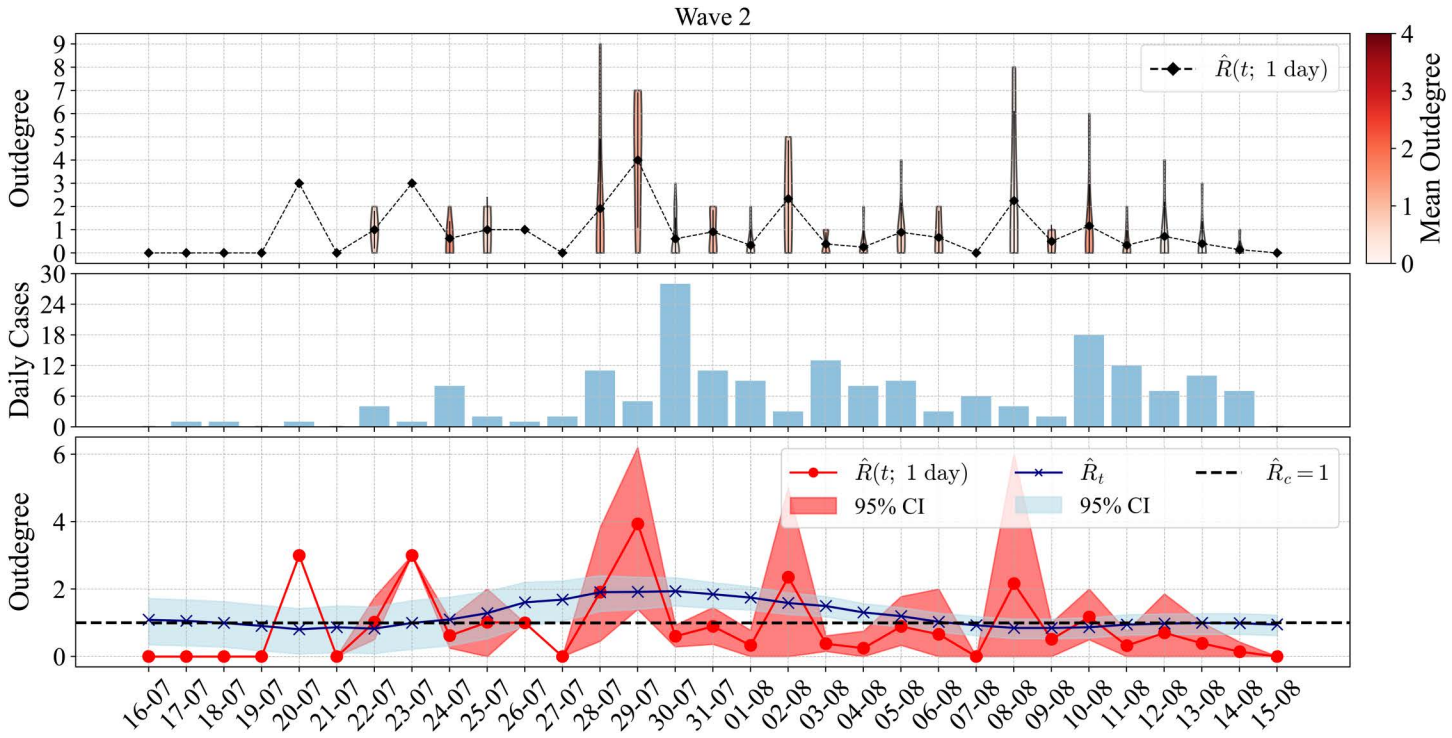


Fig 10. Daily evolution of the second wave of the COVID-19 epidemic in Cyprus. Daily out-degree distribution (top) and daily reported case counts (middle) during the second COVID-19 wave in Cyprus. Violin plots represent the distribution of secondary infections per case per day; color intensity indicates the mean outdegree. (Bottom) Evolution of the effective reproduction number $\hat{R}(t; 1\ \text{day})$ in the first wave of the COVID-19 epidemic in Cyprus. The red line shows the mean outdegree for each day derived from bootstrapped samples, with the 95% confidence intervals indicated by the red area. The blue line represents the effective reproduction number reported in [44]. In dashed black the reference line at the critical value $\hat{R}_c = 1$.

https://doi.org/10.1371/journal.pcbi.1014578.g010

reproduction numbers from [42], along with their own confidence intervals. For wave 1 we observe that for 28 out of the 31 days the value of $\hat{R}(t; 1\ \text{day})$ is lower than the $\hat{R}_t$ computed in [42]. Our estimate $\hat{R}(t; 1\ \text{day})$ becomes higher than $\hat{R}_t$ when the daily outdegree distribution has a very large variance reflecting the presence of superspreaders which are difficult to detect with ordinary compartmental models [24].

Similarly, for wave 2 we compare our estimates with the results from the report of the Cyprus Ministry of Health [44]. The methods employed by the Cyprus MoH in [44] to compute the effective reproduction number are the same as the one in [42]. Fig 10 shows that our estimate $\hat{R}(t; 1\ \text{day})$ is lower on 23 out of the 31 days of wave 2, while it is higher when there is a higher variance on the daily outdegree distribution.

Successively, we compare our measure $\hat{R}(t; 1\ \text{week})$ for the first and second wave with the reproduction number $\hat{R}_t$ computed in [42,44] employing a meta-population model of Li et al. [41] incorporating information on human movement between the five main districts of Cyprus. The comparison is displayed in Fig 11. We observe that for wave 1 the value of $\hat{R}(t; 1\ \text{week})$ is always lower than the $\hat{R}_t$ computed in [42] and follows the same trend. For wave 2 the reproduction number from [44] always falls in our confidence interval showing a good agreement with our estimate.

For waves 3 and 4, as shown in S9 Fig and S10 Fig, it is even more evident that our effective daily reproduction number $\hat{R}(t; 1\ \text{day})$, estimated from the outdegree distribution of the infection trees, consistently falls below the effective reproduction number computed with compartmental models and data from [42].

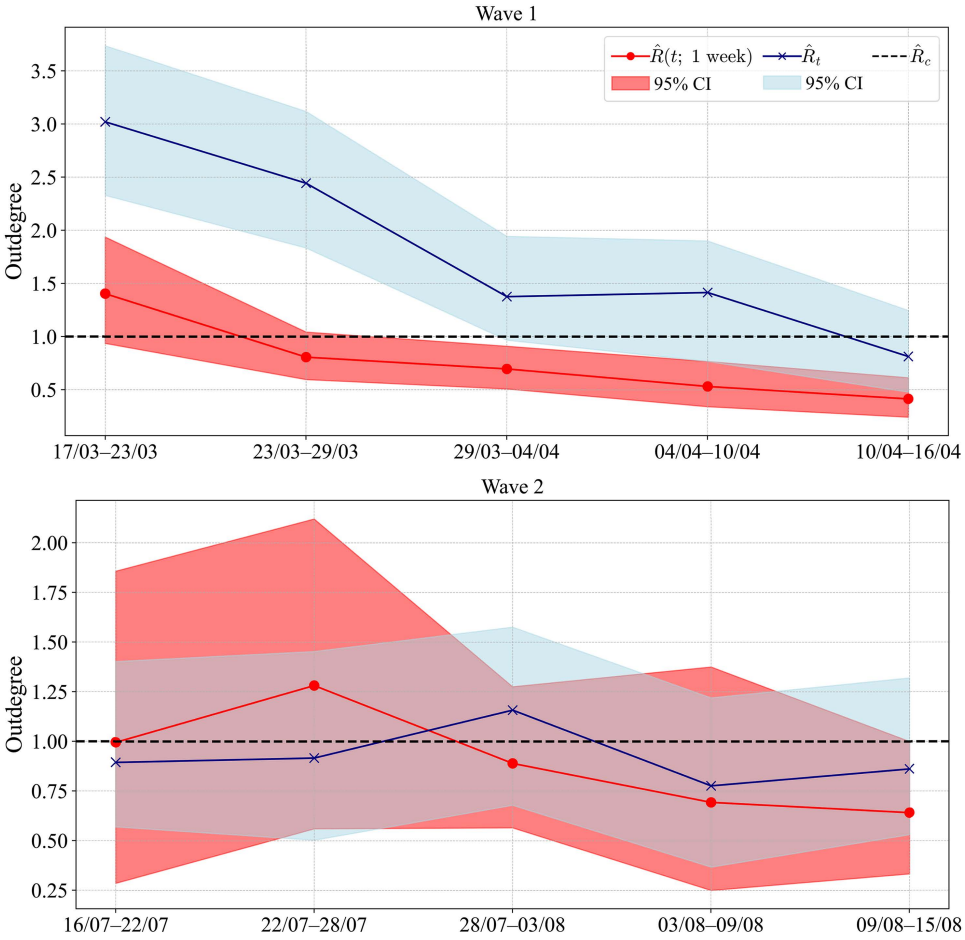


Fig 11. Evolution of the effective reproduction number $\hat{R}(t;\ 1\ \text{week})$ in the first two waves of the COVID-19 epidemic in Cyprus. The red line shows the mean outdegree for each week derived from bootstrapped samples, with the 95% confidence intervals indicated as a red area. The blue line for Wave 1 represents the effective reproduction number reported in Fig 10 of [42]. The blue line for Wave 2 displays values from [44]. In dashed black the reference line at the critical value $\hat{R}_c = 1$.

https://doi.org/10.1371/journal.pcbi.1014578.g011

We thus conclude that our estimate $\hat{R}(t;\Delta t)$ of the effective reproduction number often lower bounds measures obtained with compartmental or meta-population models [42,44]. Indeed, in the computation of the effective reproduction number $\hat{R}(t;\Delta t)$ we include all trees with $N \geq 2$ nodes, which are probably partial subtrees of larger infection trees. In small trees the majority of the nodes exhibit an average outdegree smaller than one, thereby reducing the average number of secondary cases.

Finally, a high variance of the outdegree distribution indicates the presence of nodes from larger trees. In such cases, our effective reproduction number $\hat{R}(t;\Delta t)$ increases, and leads to an overestimation of the reproduction number computed in [42,44]: nodes with very high outdegree (sometimes called superspreaders [24]) are often overlooked by mean field compartmental models.

A simple pruning technique to mitigate the underestimation of the effective reproduction number

We conclude proposing a basic procedure to mitigate the bias introduced by the children nodes in the trees: we simply prune the last level of the infection trees and we compute again $\hat{R}(t;\Delta t)$ in different time windows. Fig 12 presents $\hat{R}(t;\Delta t)$

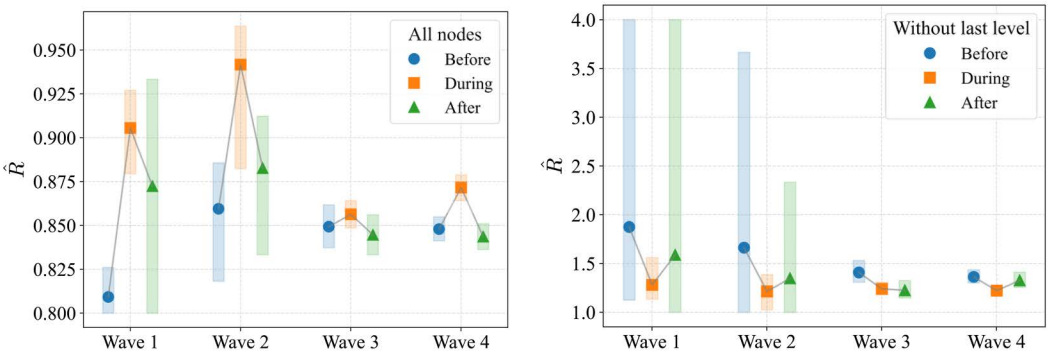


Fig 12. Average outdegree and 95% confidence intervals for the infection trees belonging to different periods around the peak of the four COVID-19 waves in Cyprus. The trees are divided into three groups: trees whose nodes were only tested positive before the peak (Before), trees whose nodes were tested positive before and after the peak (During), and trees whose nodes were only tested positive after the peak (After). On the left the plot when no pruning is applied. On the right the estimate where the last level of the infection trees is pruned.

https://doi.org/10.1371/journal.pcbi.1014578.g012

computed in three different time windows before, after or during the peak of the wave (maximum number of confirmed daily cases). On the left the effective reproduction number has been computed retaining the full structure of the trees, on the right removing the last level of any tree. The trees are divided into three groups: trees whose nodes were tested positive only before the peak (Before), trees whose nodes were tested positive before and after the peak (During), and trees whose nodes were only tested positive after the peak (After).

From the results in Fig 12, the estimate without the last level (right figure) is the one which reasonably reflects the expected evolution of the epidemic wave. Without the last level we observe a decrease in the effective reproduction number from before to after the peak of each wave. Retaining all the nodes shows instead smaller relative variations before and after the peak or even an increase in the effective reproduction number, a behavior that is inconsistent with the observed epidemic trajectory. Similar conclusions can be drawn from S11 Fig for the daily effective reproduction number $\hat{R}(t; 1\text{ day})$. The infection trees without the last level provide a more conservative estimate of the effective reproduction number, avoiding the underestimation introduced by leaf nodes.

Discussion

In this work, we extract from the topology of real infection trees a non-Markovianity measure α, which is the shape parameter of a Weibull (6) distribution. The hopcount distribution alone reveals that infection trees of varying sizes can be described as realizations of a spreading process with non-exponential infection times characterized by a Weibull distribution with shape parameter $\alpha > 1$. Fig 8 illustrates the results for different waves and different tree sizes, highlighting the general non-Markovianity of the underlying spreading process. In our model, the infection time acts as an *effective transmission time* that inherently captures multiple, often unobservable, delays, ranging from true biological latency to human behavioral factors. Additionally, our non-exponential infection time corresponds to an effective non-exponential generation interval [37], in line with previous evidence from COVID-19 [17,18].

Generalized epidemic models [46] can be created by extending the basic SI process with additional compartments (e.g., recovered, exposed) and are hence able to mimic non-exponential infection times via a sum of independent exponential delays. Although such generalized models may provide a valid statistical description of the data (S7 Fig and S1 Table), the underlying Markovian assumptions can lead to fitting parameters with values that contradict established timelines measured in COVID-19 (e.g., incubation rate much larger than the transmission rate S1 Table). Furthermore, generalized epidemic models require additional parameters that are quite difficult to measure, in contrast to our

phenomenological approach, that systematically absorbs all the complexity into a single parameter α. Our approach is further validated when a recovery compartment is introduced: S8 Fig shows that a non-Markovian SIR process still necessitates an $\alpha > 1$ to match the empirical data, confirming that the deviations from the Markovianity are not an artifact of our SI modeling assumption. The well-established Markovian case $\alpha = 1$, that produces nearly all currently used mean-field epidemic models, may thus provide a safe upper bound for the nodal infection probabilities in the contact graph and an upper bound for the basic reproduction number [37].

Our computations are based on the assumption that each observed infection tree of size N is an independent realization of a non-Markovian SI process on the complete graph K_N. The assumption of independence between the infection trees is justifiable in large populations, in particular when the global contact network is unknown. By treating the trees as separate and independent, we neglect possible, but hard-to-measure dependencies between trees. The resulting estimate of α per tree size is derived from localized transmission events.

Each infection process in the network is assumed to be independent from the others and the infection time is Weibull distributed with shape parameter α and rate parameter $\beta = 1$. The choice of the value for the rate parameter β does not affect the SI process [28, Sec. 16.3]: when the infection times are i.i.d random variables, a global rescaling of all link infection times does not change the infection tree structure.

The observed infection trees are usually small and do not contain information about the curings, which makes SI modeling meaningful. The complete graph assumption and the absence of curing processes overestimate the viral force of a single node to infect its neighbors. Adding recoveries in the SI process often requires a higher value of α to obtain a hopcount distribution closer to the empirical one, as shown in S8 Fig, meaning that our estimates of the non-Markovian parameter α are safely larger than 1.

Although contact tracing investigations in Cyprus were conducted thoroughly, it was nearly impossible to trace all infections and construct a single connected infection tree per each wave. As a result, the observed infection trees are sparse and affected by under-reporting or incomplete case detection [7,33,47,48]. Such limitations may directly affect our estimates: if branches from the root node are missing, the computed α value is underestimated, while if branches are missing from the leaf nodes, α is overestimated. Nonetheless, the key qualitative conclusion $\alpha > 1$ is observed consistently across different waves and multiple tree sizes.

Our estimates of the reproduction number tend to produce lower values compared to previous studies [42,44], except when the daily outdegree distribution variance is very large. To mitigate the effect of the leaf nodes, a simple pruning that removes the last level of the infection trees results in an effective reproduction number estimate which follows more accurately the evolution of the epidemic, as shown in Fig 12.

Conclusion

We show that real-world epidemics are often characterized by non-exponential infection processes and, consequently, by non-Markovian dynamics. In this study, we quantify the non-Markovianity by the parameter α. Empirical data consistently points to a value of $\alpha > 1$, which indicates that the infection process likely exhibits non-Markovian characteristics. Our study statistically rejects the Markovian hypothesis ($\alpha = 1$), based solely on the topology of real contact-tracing data.

Additionally, the analysis of infection trees offers an independent and complementary approach for estimating the effective reproduction number, a key epidemiological metric. Finally, our study demonstrates the value of "infection tree construction" and stimulates the design of digital apps that are able to automatically reconstruct the underlying transmission network [49,50].

Supporting information

S1 File. Supplementary material, which includes S1-S5 Appendixes, S1-S11 Figs, and S1 Table.
(PDF)

Author contributions

Conceptualization: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Piet Van Mieghem.

Data curation: Pavlos Alexandros Dimitriou, Valentinos Silvestros, Elisavet Constantinou, Costas Pitris, Panayiotis Kolios.

Formal analysis: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Piet Van Mieghem.

Funding acquisition: Costas Pitris, Panayiotis Kolios, Piet Van Mieghem.

Methodology: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Brian L. Chang, Piet Van Mieghem.

Project administration: Piet Van Mieghem.

Resources: Piet Van Mieghem.

Software: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Brian L. Chang.

Supervision: Piet Van Mieghem.

Validation: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Brian L. Chang, Piet Van Mieghem.

Visualization: Pavlos Alexandros Dimitriou, Matteo D'Alessandro.

Writing – original draft: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Piet Van Mieghem.

Writing – review & editing: Pavlos Alexandros Dimitriou, Matteo D'Alessandro, Brian L. Chang, Piet Van Mieghem.

References

1. World Health Organization. COVID-19 Cases, World. World Health Organization; 2025. [cited 2025 Feb 14]. Available from: https://data.who.int/dashboards/covid19/cases
2. Liu Y, Gu Z, Liu J. Uncovering transmission patterns of COVID-19 outbreaks: a region-wide comprehensive retrospective study in Hong Kong. EClinicalMedicine. 2021;36:100929. https://doi.org/10.1016/j.eclinm.2021.100929 PMID: 34124628
3. Liu Y, Gu Z, Xia S, Shi B, Zhou X-N, Shi Y, et al. What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. EClinicalMedicine. 2020;22:100354. https://doi.org/10.1016/j.eclinm.2020.100354 PMID: 32313879
4. White LF, Moser CB, Thompson RN, Pagano M. Statistical estimation of the reproductive number from case notification data. Am J Epidemiol. 2021;190(4):611–20. https://doi.org/10.1093/aje/kwaa211 PMID: 33034345
5. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004;160(6):509–16. https://doi.org/10.1093/aje/kwh255 PMID: 15353409
6. Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, et al. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proc Biol Sci. 2003;270(1511):121–7. https://doi.org/10.1098/rspb.2002.2191 PMID: 12590749
7. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021;371(6526):eabe2424. https://doi.org/10.1126/science.abe2424 PMID: 33234698
8. Xiang Y, Jia Y, Chen L, Guo L, Shu B, Long E. COVID-19 epidemic prediction and the impact of public health interventions: a review of COVID-19 epidemic models. Infect Dis Model. 2021;6:324–42. https://doi.org/10.1016/j.idm.2021.01.001 PMID: 33437897
9. Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P. Network epidemiological analysis of COVID-19 transmission patterns by age, occupation and residence across four waves in Cyprus. Sci Rep. 2025;15(1):28416. https://doi.org/10.1038/s41598-025-14267-7 PMID: 40760155
10. Taube JC, Miller PB, Drake JM. An open-access database of infectious disease transmission trees to explore superspreader epidemiology. PLoS Biol. 2022;20(6):e3001685. https://doi.org/10.1371/journal.pbio.3001685 PMID: 35731837
11. Soetens L, Klinkenberg D, Swaan C, Hahné S, Wallinga J. Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak. Epidemiology. 2018;29(2):230–6. https://doi.org/10.1097/EDE.0000000000000776 PMID: 29087987
12. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. Rev Mod Phys. 2015;87(3):925–79. https://doi.org/10.1103/revmodphys.87.925
13. Van Mieghem P, Blenn N, Doerr C. Lognormal distribution in the Digg online social network. Eur Phys J B. 2011;83(2):251–61. https://doi.org/10.1140/epjb/e2011-20124-0
14. Doerr C, Blenn N, Van Mieghem P. Lognormal infection times of online information spread. PLoS One. 2013;8(5):e64349. https://doi.org/10.1371/journal.pone.0064349 PMID: 23700473

15. Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, et al. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. Am J Epidemiol. 2008;167(7):775–85. https://doi.org/10.1093/aje/kwm375 PMID: 18230677
16. Cowling BJ, Fang VJ, Riley S, Malik Peiris JS, Leung GM. Estimation of the serial interval of influenza. Epidemiology. 2009;20(3):344–7. https://doi.org/10.1097/EDE.0b013e31819d1092 PMID: 19279492
17. Kim D, Ali ST, Kim S, Jo J, Lim J-S, Lee S, et al. Estimation of serial interval and reproduction number to quantify the transmissibility of SARS-CoV-2 omicron variant in South Korea. Viruses. 2022;14(3):533. https://doi.org/10.3390/v14030533 PMID: 35336939
18. Backer JA, Eggink D, Andeweg SP, Veldhuijzen IK, van Maarseveen N, Vermaas K. Shorter serial intervals in SARS-CoV-2 cases with Omicron BA.1 variant compared with Delta variant, the Netherlands, 13 to 26 December 2021. Eurosurveillance. 2022;27(6). https://doi.org/10.2807/1560-7917.ES.2022.27.6.2200042
19. Van Mieghem P, van de Bovenkamp R. Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. Phys Rev Lett. 2013;110(10):108701. https://doi.org/10.1103/PhysRevLett.110.108701 PMID: 23521310
20. Cator E, van de Bovenkamp R, Van Mieghem P. Susceptible-infected-susceptible epidemics on networks with general infection and cure times. Phys Rev E Stat Nonlin Soft Matter Phys. 2013;87(6):062816. https://doi.org/10.1103/PhysRevE.87.062816 PMID: 23848738
21. Liu Q, Van Mieghem P. Burst of virus infection and a possibly largest epidemic threshold of non-Markovian susceptible-infected-susceptible processes on networks. Phys Rev E. 2018;97(2–1):022309. https://doi.org/10.1103/PhysRevE.97.022309 PMID: 29548175
22. Van Mieghem P, Liu Q. Explicit non-Markovian susceptible-infected-susceptible mean-field epidemic threshold for Weibull and Gamma infections but Poisson curings. Phys Rev E. 2019;100(2–1):022317. https://doi.org/10.1103/PhysRevE.100.022317 PMID: 31574702
23. Di Lauro F, KhudaBukhsh WR, Kiss IZ, Kenah E, Jensen M, Rempała GA. Dynamic survival analysis for non-Markovian epidemic models. J R Soc Interface. 2022;19(191):20220124. https://doi.org/10.1098/rsif.2022.0124 PMID: 35642427
24. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–9. https://doi.org/10.1038/nature04153 PMID: 16292310
25. Plazzotta G, Kwan C, Boyd M, Colijn C. Effects of memory on the shapes of simple outbreak trees. Sci Rep. 2016;6:21159. https://doi.org/10.1038/srep21159 PMID: 26888437
26. Van Mieghem P, Liu Q. Explicit non-Markovian susceptible-infected-susceptible mean-field epidemic threshold for Weibull and Gamma infections but Poisson curings. Phys Rev E. 2019;100(2–1):022317. https://doi.org/10.1103/PhysRevE.100.022317 PMID: 31574702
27. Aldous D. Recursive self-similarity for random trees, random triangulations and Brownian excursion. Ann Probab. 1994;22(2). https://doi.org/10.1214/aop/1176988720
28. Van Mieghem P. Performance analysis of complex networks and systems. Cambridge University Press; 2014. https://doi.org/10.1017/CBO9781107415874
29. Van Mieghem P, Hooghiemstra G, van der Hofstad R. A scaling law for the hopcount in internet. Delft University of Technology; 2000.
30. Van Mieghem P, Hooghiemstra G, van der Hofstad RW. Modeling the AS hopcount in Internet. Delft University of Technology; 2002.
31. Begtasevic F, Van Mieghem P. Measurements of the Hopcount in Internet. In: Proceedings of Workshop on Passive and Active Measurement (PAM). Citeseer; 2001. pp. 183–90.
32. Anastasiadou MN, Isaia P, Kolios P, Eliades DG, Laoudias C. Leveraging ICTs for Effective COVID-19 Pandemics Management: Insights from a Health Information System Implementation in Cyprus. In: 2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW). IEEE; 2024. pp. 84–91.
33. Nagarajan K, Muniyandi M, Palani B, Sellappan S. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. BMC Med Res Methodol. 2020;20(1):233. https://doi.org/10.1186/s12874-020-01119-3 PMID: 32942988
34. Persoons R, Van Mieghem P. Finding patient zero in susceptible-infectious-susceptible epidemic processes. Phys Rev E. 2024;110(4–1):044308. https://doi.org/10.1103/PhysRevE.110.044308 PMID: 39562901
35. Abramowitz M, Stegun IA. Handbook of mathematical tables. Dover Publications; 1968.
36. Bhamidi S, Van Der Hofstad R. Weak disorder asymptotics in the stochastic mean-field model of distance. Ann Appl Probab. 2012;22(1):29–69. https://doi.org/10.1214/10-AAP753
37. Chang BL, Van Mieghem P. Classical mean-field epidemic theory upper-bounds real disease spread. Commun Phys. 2026. https://doi.org/10.1038/s42005-026-02723-3
38. Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. J Math Biol. 1990;28(4):365–82. https://doi.org/10.1007/BF00178324 PMID: 2117040
39. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc London Series A, Contain Pap Math Phys Character. 1927;115(772):700–21. https://doi.org/10.1098/rspa.1927.0118
40. van den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math Biosci. 2002;180:29–48. https://doi.org/10.1016/s0025-5564(02)00108-6 PMID: 12387915
41. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368(6490):489–93. https://doi.org/10.1126/science.abb3221 PMID: 32179701

42. Agapiou S, Anastasiou A, Baxevani A, Nicolaides C, Hadjigeorgiou G, Christofides T, et al. Modeling the first wave of Covid-19 pandemic in the Republic of Cyprus. Sci Rep. 2021;11(1):7342. https://doi.org/10.1038/s41598-021-86606-3 PMID: 33795723
43. Dijkstra EW. A note on two problems in connexion with graphs. In: Apt KR, Hoare T, editors. Edsger Wybe Dijkstra: His Life, Work, and Legacy. ACM Books; 2022. pp. 2879–90. https://doi.org/10.1145/3544585.3544600
44. Office CGP. National Report on COVID-19 in Cyprus up to 22/09/2020. 2020. [cited 2025 Apr 7]. Available from: https://www.pio.gov.cy/coronavirus/pdf/ep22092020.pdf
45. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv preprint arXiv:200206563. 2020. https://doi.org/10.48550/arXiv.2002.06563
46. Darabi Sahneh F, Scoglio C, Van Mieghem P. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. IEEE/ACM Trans Networking. 2013;21(5):1609–20. https://doi.org/10.1109/tnet.2013.2239658
47. Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P. Reconstructing COVID-19 Infection Networks: A Link Prediction Approach to Connecting Orphan Cases. In: Cherifi H, Rocha LM, Cherifi C, Ertem MZ, editors. Complex Networks & Their Applications XIV. Cham: Springer Nature Switzerland; 2026. pp. 29–39. https://doi.org/10.1007/978-3-032-16649-4_3
48. Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P. Predicting Missing Links in COVID-19 Infection Networks. 2026. https://doi.org/10.1371/journal.pcsy.0000124
49. Ahmed N, Michelin RA, Xue W, Ruj S, Malaney R, Kanhere SS, et al. A Survey of COVID-19 Contact Tracing Apps. IEEE Access. 2020;8:134577–601. https://doi.org/10.1109/access.2020.3010226
50. Prasse B, Van Mieghem P. Mobile smartphone tracing can detect almost all SARS-CoV-2 infections. arXiv preprint arXiv:200614285. 2020. https://doi.org/10.48550/arXiv.2006.14285