Network science inspires novel tree shape statistics

Leonid Chindelevitch
MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London
London, United Kingdom, W2 1NY

Maryam Hayati
School of Computing Science, Simon Fraser University
Burnaby, BC, Canada, V5A 1S6

Art F. Y. Poon
Department of Pathology & Laboratory Medicine, University of Western Ontario
London, ON, Canada, N6A 5C1

Caroline Colijn
Department of Mathematics, Simon Fraser University
Burnaby, BC, Canada, V5A 1S6

February 1, 2021

1 Abstract

The shape of phylogenetic trees can be used to gain evolutionary insights. A tree's shape specifies the connectivity of a tree, while its branch lengths reflect either the time or genetic distance between branching events; well-known measures of tree shape include the Colless and Sackin imbalance, which describe the asymmetry of a tree. In other contexts, network science has become an important paradigm for describing structural features of networks and using them to understand complex systems, ranging from protein interactions to social systems. Network science is thus a potential source of many novel ways to characterize tree shape, as trees are also networks. Here, we tailor tools from network science, including diameter, average path length, and betweenness, closeness, and eigenvector centrality, to summarize phylogenetic tree shapes. We thereby propose tree shape summaries that are complementary to both asymmetry and the frequencies of small configurations. These new statistics can be computed in linear time and scale well to describe the shapes of large trees. We apply these statistics, alongside some conventional tree statistics, to phylogenetic trees from three very different viruses (HIV, dengue fever and measles), from the same virus in different epidemiological scenarios (influenza A and HIV) and from simulation models known to produce trees with different shapes. Using mutual information and supervised learning algorithms, we find that the statistics adapted from network science perform as well as or better than conventional statistics. We describe their distributions and prove some basic results about their extreme values in a tree. We conclude that network science-based tree shape summaries are a promising addition to the toolkit of tree shape features. All our shape summaries, as well as functions to select the most discriminating ones for two sets of trees, are freely available as an R package at `http://github.com/Leonardini/treeCentrality`.

2 Introduction

Molecular data describing the evolution, variation and diversity of organisms over time is more widely available than ever before due to rapid improvements in sequencing technology. Using these data to infer the underlying evolutionary process is a key ongoing challenge in many areas of biology. In particular, in infectious disease, it is crucial to understand pathogen adaptation: despite improvements in sanitation and vaccination and the development of antibiotics, infectious pathogens continue to emerge from zoonotic infections and to adapt to human immune responses, vaccines, and antimicrobials. Next-generation sequencing has afforded unprecedented opportunities to generate pathogen genome sequences in a highly scalable manner, and theoretical tools have been developed to interrogate these data, largely through reconstructed phylogenetic trees.

There has been considerable interest over the years in comparing the shapes of phylogenetic trees in order to understand evolutionary processes [64, 63, 24, 58, 31, 47, 6, 71]. A tree's shape specifies its connectivity structure. The lengths of its branches typically reflect either the time or genetic distance between branching events. Following the observation that reconstructed evolutionary trees are more asymmetric than random models predict [2], there have been efforts to summarise tree asymmetry in trees reconstructed from data and relate it to predicted asymmetry in evolutionary and ecological models [31, 19, 3, 55, 66, 47, 6, 58, 38]. There is also interest in establishing whether taxa from two phylogenies might correspond to each other, for example in the context of parasites and hosts or fossils of different origins [22], and in comparing simulated trees with trees from data in epidemiology, for example using Approximate Bayesian Computation [56, 62]. These applications require quantitative tools to compare phylogenetic trees with different taxa, and they require summary features that are informative of the evolution or epidemiology being studied.

Summaries of tree topology have often focused on either asymmetry, or the frequency of various configurations such as cherries or ladders [60]. Well-known measures of asymmetry include the Colless and Sackin imbalance [12, 61]. Asymmetry measures tend to be correlated with each other, and do not fully capture the shape of a tree [22, 40], leading to an interest in exploring other statistics, comparisons tools and metrics for this task [27, 57, 36, 22, 11, 62]. Some metric approaches directly find a distance (or similarity) measure between unlabelled phylogenies; others seek an optimal labelling for two unlabelled trees and use metrics for labelled trees, but this is not feasible for large trees. Metric approaches also do not lend themselves to interpretable descriptions of trees that can easily be connected with generative models of evolution or epidemiology.

In one particularly important application of tree shape statistics, namely, Approximate Bayesian Computation, one seeks simulation parameters that produce trees that are similar enough to data-derived trees [14]. This leaves a need for new statistics to quickly provide a good summary of tree topology. Matsen used optimization over binary recursive tree shape statistics to find statistics that distinguish between sets of trees [41]. This set the stage for broadening the set of quantities used to describe trees, and gives rise to natural questions like why a particular recursive statistic separates two groups, whether it only happens to separate those specific trees or acts as a good way to distinguish between trees that are meaningfully similar to those in the two groups. The binary recursive class may furthermore exclude features that capture global information about a tree, such as features derived from its eigenvalues when viewed as a graph, known as spectral features. As the size of datasets and range of applications increases, it is reasonable to assume that inference will be improved by expanding upon the tools available to summarize tree topologies [41, 62]. In applications like Approximate Bayesian Computation, as well as other related attempts at summarizing tree shapes, one does not necessarily seek to relate summaries to evolutionary mechanisms. Rather, the key objective is to distinguish between trees in different categories or scenarios in an efficient way by computing simple statistics.

Network science has become an important paradigm for describing structural (topological)

features of networks and using them to understand complex systems, ranging from protein interactions to social systems [35, 49]. Network science is thus a potential source of many novel ways to characterize tree topology. A phylogeny can be interpreted as a simple type of network or graph – specifically, a connected acyclic undirected graph. Thinking of a tree as a graph can lead to a matrix representation (such as the adjacency matrix) in which both rows and columns correspond to its internal and terminal nodes. This makes available the large assortment of techniques for studying the properties of graphs based on their matrix representation [25]. The *spectra* of graph matrices have recently been used to compare tree topologies [36]. However, the use of classical spectra is not likely to uniquely define trees [43, 42], the computation does not scale well with tree size, and describing the relationship between evolutionary or epidemiological parameters and graph spectra may be extremely difficult due to the algebraic rather than combinatorial nature of the latter.

Network science offers many other network features that can be adapted to describe shape. Here, we tailor tools from network science to summarize phylogenetic tree topologies. We thereby develop tree topology summaries that are complementary to both asymmetry and the frequencies of small configurations. These new statistics are fast to compute and will scale well to describe the topologies of large trees. They can additionally be easily adapted to take branch lengths into account. We illustrate how these statistics vary, alongside some conventional tree statistics, over three very different viruses (HIV, dengue fever and measles), and over the same virus in different epidemiological scenarios (influenza A over a 2-year period in the USA, over a 5-year period globally, and over a longer period globally; HIV in three epidemiological contexts, namely a concentrated epidemic in men who have sex with men, a generalised African HIV epidemic in a village setting and HIV-1 from a national-level survey (see Methods)). We also use the statistics to compare simulated trees from different models. Finally, we use classification with and without the network-based statistics to distinguish trees in different settings. We find that the network-science statistics improve the classification performance and are consistently assigned a high importance measure in classification algorithms.

3 Methods

3.1 Definitions related to graphs

In this subsection we introduce some definitions from graph theory, used below. They are based on the textbooks by Bollobás [7] as well as Godsil and Royle [20].

A *network* or *graph* G consists of a finite set of *nodes* V and a finite set of *edges* E whose elements are unordered pairs of distinct nodes. The edge $\{u, v\}$, denoted uv for simplicity, *joins* nodes u and v, and if $uv \in E$, the nodes u and v are *adjacent*, which we denote by $u \sim v$.

A graph is *weighted* if there is a function $w : E \to \mathbb{R}$ that assigns a *weight* $w(uv)$ to each edge $uv \in E$. The weight function extends to all of $V \times V$ by setting $w(uv) = 0$ if $u = v$ or $uv \notin E$.

A *path* P in a graph is a sequence of nodes $v_0, v_1, \ldots, v_k$ such that $v_i v_{i+1} \in E$ for each $0 \leq i \leq k-1$. The nodes v_0 and v_k are the *endvertices*. In the unweighted case, k is the *length* of P. In the weighted case, the weighted length of the path P is measured by the sum of the weights of the edges it contains, i.e. $\sum_{i=0}^{k-1} w(v_i v_{i+1})$.

The path P corresponding to the sequence $v_0, v_1, \ldots, v_k$ is a *simple path* if $v_i \neq v_j$ for any $0 \leq i < j \leq k$. If the endvertices are identical, i.e. $v_0 = v_k$, then P is a *cycle*; if the endvertices are the only identical nodes on the cycle, it is a *simple cycle*. A simple cycle of length 2 is just an edge traversed twice; a graph with no simple cycle of length greater than 2 is called *acyclic*.

In both the unweighted and the weighted cases, the minimum (weighted) length of a path with endvertices u and v is called the (weighted) *distance* between u and v and denoted $d(u, v)$. The (weighted) *diameter* of the graph is the largest distance between two distinct nodes.

Being the endvertices of some path of G defines an equivalence relation on the set of nodes, and its equivalence classes are called the *connected components* (or simply *components*) of G. A graph is *connected* if it contains only one component. A node whose deletion increases the number of components is a *cutvertex*; an edge whose deletion increases this number is a *bridge*.

The *neighborhood* $\Gamma(u)$ of a node $u \in G$ is the set of nodes adjacent to it in G. The *degree* of u is $d(u) = |\Gamma(u)|$. The *degree sequence* of G is the vector $\mathbf{d} = (d(v_1) \dots d(v_n))$ containing the degree of all its nodes in some order. In a weighted graph, the *weighted degree* of u is $d_w(u) = \sum_v w(uv)$, and the *weighted degree sequence* is the vector $\mathbf{d_w} = (d_w(v_1) \dots d_w(v_n))$. The *density* of G is the ratio $\rho := |E|/|V|$; if G is unweighted, ρ equals half the average degree.

The *clustering coefficient* of a node u in a graph is defined as $\frac{|E(\Gamma(u))|}{\binom{d(u)}{2}}$, where $E(\Gamma(u))$ is the set of pairs $v, v' \in \Gamma(u)$ such that $vv' \in E$. This coefficient measures the fraction of mutually adjacent nodes among the nodes adjacent to u, and varies from 0 to 1. The *assortativity coefficient* of G is the Pearson correlation coefficient of degree between pairs of adjacent nodes.

Given a numbering of the nodes, $V = \{1, 2, \dots, n\}$, a graph can be represented by its *adjacency matrix* $A \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$ if $ij \in E$ and $A_{ij} = 0$ otherwise. If the graph is weighted, its *weighted adjacency matrix* has entries $A_{ij} = w(ij)$.

A graph can also be represented by its *Laplacian matrix* $L \in \mathbb{R}^{n \times n}$, defined by $L_{ii} = d(i)$ for all i, $L_{ij} = -1$ if $ij \in E$ and $L_{ij} = 0$ otherwise. The *weighted Laplacian matrix* has entries $L_{ii} = d_w(i)$ for all i and $L_{ij} = -w(ij)$ if $i \neq j$. Thus, in both cases we have $L = D - A$, where D is the diagonal matrix with the (weighted) degree sequence on the diagonal.

The *normalized Laplacian* [9] is a variant of the Laplacian which is normalized by the degree of each node, and is defined as $\mathcal{L} = D^{-1/2} L D^{-1/2}$. Its entries are $\mathcal{L}_{ii} = 1$ for all i and $\mathcal{L}_{ij} = -\frac{w(ij)}{\sqrt{d_i d_j}}$ for $i \neq j$ in the weighted case or $\mathcal{L}_{ij} = -\frac{1}{\sqrt{d_i d_j}}$ in the regular case.

In addition, a graph can also be represented by its *distance matrix* $Q \in \mathbb{R}^{n \times n}$, with $Q_{ij} = d(i, j)$, the (weighted or unweighted) graph distance between i and j, and in particular, $Q_{ii} = 0$. As proposed in [36], it is also possible to apply the (regular or normalized) Laplacian transformation to the distance matrix (by treating the distance matrix as a weighted adjacency matrix) to obtain alternative matrix representations of a graph.

Widely studied graph properties fall into two broad classes. The first are *shape properties*, concerned with various connectivity measures of the graph, such as the numbers of distinct paths connecting pairs of nodes and the average or longest distance between a pair of nodes. Some of these properties do not depend on the edge weights of the graph, in which case we refer to them as *topology properties*; we use the term *shape property* as a broad term that includes both topology properties and properties that take edge weights into account. The second are *spectral properties*, concerned with the eigenvalues (and occasionally eigenvectors) of matrices representing the graph, such as the adjacency, Laplacian, and normalized Laplacian.

3.2 Definitions related to trees

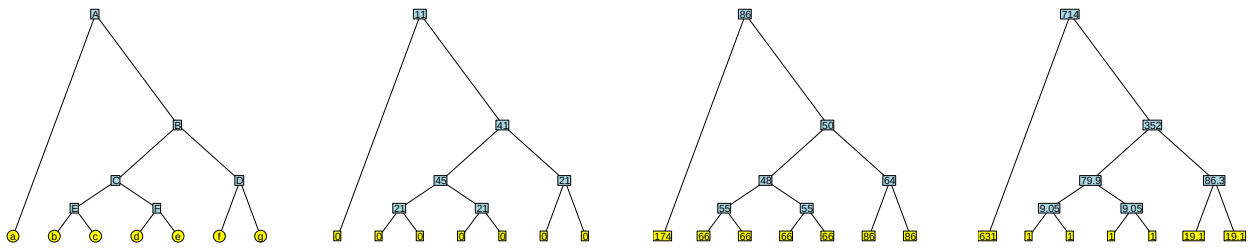
A tree is a connected acyclic graph. It can be shown that a tree with n nodes has $n-1$ edges, and that there is a unique simple path between any two nodes in a tree. The nodes of degree one are called *tips* (or *leaves*); the remaining nodes are called *internal*. The weight $w(e)$ of an edge e of a tree is also called its *branch length*; we assume that a tree has all positive branch lengths.

A tree T is *rooted* if it contains a distinguished node r, called the *root*. If v is any node other than the root r in a rooted tree, the node u immediately preceding v on the unique simple path from v to r is called the *parent* of v in T, and any node z on this unique simple path is called an *ancestor* of v in T. In particular, the root r is an ancestor of every node in T other than itself. The number of edges on the unique shortest path from a node v to the root r is called its *depth*.

Conversely, in this situation, v is called a *child* of u and a *descendant* of z. A rooted tree T is *binary* if every internal node has exactly two children. In a rooted tree T, the set of all descendants of a node u including u itself forms a *clade*; we say that u *subtends* this clade. The *subtree* rooted at u, denoted T_u, is the clade that u subtends together with all the edges of T connecting its elements. If a node u has two children v and w, we may arbitrarily refer to them as the left and right child of u, and call T_v and T_w the left and right subtree of u, respectively.

A tree T with root node r can be *rerooted* at an internal node $s \neq r$, by simply designating s rather than r as the root. This changes only the ancestor-descendant relationships, but not the topology of the tree. Note that if T is a binary tree with root r, T rerooted at $s \neq r$ will not be binary because s will have three children while r will have only one child after rerooting. However, this situation is often resolved by specifying a branching order, i.e. creating a new node t that is the parent of two children of the new root s, and adding an $s - t$ edge of length 0. An analogous procedure, called "multichotomy resolution", may be iteratively applied to any other internal nodes with a degree greater than 2 to turn any non-binary tree into a binary tree.

A molecular *phylogeny*, or *phylogenetic tree*, is a binary rooted tree whose tips correspond to genetic or genomic sequences, and whose internal nodes represent their inferred common ancestors. A phylogeny therefore represents the ancestral relationships among a set of genomes. The shape of a phylogeny tells the story of an evolutionary history going back through time to the most recent common ancestor at the root of the tree (see Figure 1 (a)).



(a) (b) (c) (d)

Figure 1: (a) An example tree. In the epidemiological context, tips $(a - g)$ would correspond to pathogen sequences and internal nodes $(A - F)$ to their inferred common ancestors. Node D subtends a "cherry" configuration, and node C subtends two cherries (a "double cherry"). The heights of the internal nodes are 1 (E, F), 2 (C, D), 4 (B) and 8 (A), so the diameter is 16 and the Wiener index is 484, for a mean path length of 6.21.
(b) Same tree with betweenness centrality values at each node (note that branch lengths do not change them). The tree has betwenness centrality 45.
(c) Same tree with farness (reciprocal of closeness centrality) values at each node. The tree has closeness centrality 1/48.
(d) Same tree with eigenvector centrality values (scaled to have a minimum of 1) at each node, rounded to 3 significant figures. Here, the leading eigenvalue is $\lambda = 9.05$. By our definition, the tree has eigenvector centrality $714/1023 = 0.698$ (here, 1023^2 is the sum of the squared values).

3.3 Data and simulations

HIV/Dengue/Measles We obtained Newick tree strings corresponding to phylogenies inferred from human and zoonotic RNA viruses from a previous study. Specifically, we retrieved tree strings reconstructed from genetic sequences of HIV-1 subtype B, Measles virus, and Dengue virus serotype 4. The HIV sequence data (corresponding to the gene encoding the Nef pro-

tein) were obtained from the LANL HIV Sequence database [18], through the web site at *http://www.hiv.lanl.gov*, and screened for recombinants using the SCUEAL algorithm [32]. The remaining virus sequences were obtained from GenBank [4].

Phylogenies were reconstructed from random samples of 100 sequences by maximum likelihood using RAxML [65] under a general time-reversible model of nucleotide substitution and rate variation among sites approximated by the GTRCAT model. HIV-1 subtype B phylogenies were rooted using a subtype D sequence as an outgroup. Dengue virus serotype 4 phylogenies were rooted on an outgroup sequence isolated in the Philippines in 1956. Finally, measles virus phylogenies were rooted using a genotype D6 sequence as the outgroup. The GenBank [4] accession numbers for all outgroups can be found in Table S1 in the **Supplementary Materials**.

HIV in three settings We obtained HIV-1 sequence data from three published studies. The Wolf et al. [70] data set corresponds to samples from a concentrated epidemic of HIV-1 subtype B in populations of predominantly men who have sex with men in Seattle, USA. The Novitsky et al. [53] data set corresponds to samples from a generalized epidemic of HIV-1 subtype C infections in Mochudi, Botswana, a village with an estimated HIV-1 prevalence of about 20% in the adult population. Similarly, the Hunt et al. [28] data set represents samples from a national survey of the generalized epidemic of HIV-1 subtype C in South Africa. Thus, these studies represent a range of geographic scales and epidemiological contexts.

Influenza in three settings We compared the topologies of a single virus (human influenza A) sampled to reflect different epidemiology: (1) samples over a five-year period; (2) global samples over a 12-year period and (3) samples from 2012-2013 from the USA only. We downloaded full-length hemagglutinin (HA) sequences of human H3N2 flu from NCBI and aligned sequences from each group with MAFFT [30]. For each sample, we chose 120 sequences uniformly at random from the alignment, and inferred a tree with these sequences as tips using IQtree [67] with the pll (phylogenetic likelihood library) option [17] and the GTR+G model. Using the date information from NCBI, we rooted the trees using the root to tip (*rtt*) function in the *ape* package [54] in R [59].

Simulated tree models We simulated trees from four random processes: a Yule process (pure birth trees), a "biased" model of Kirkpatric and Slatkin [31] in which speciation rates are unevenly assigned to a node's descendants with a bias (here 0.3), and two constant rate birth-death processes, with the basic reproduction number (mean of the offspring distribution) equal to 1.5 and 3. We created sets of 100 trees with 100 tips and separately with 300 tips. The `apTreeshape` package was used to simulate the Yule and biased models; tree shapes were converted to phylogenetic trees using the `as.phylo` function. We used the `TreeSim` package for the birth-death models. Because in `sim.bd.taxa` (in `TreeSim`) the simulations are conditioned on having a fixed number of *extant* tips, we created trees with 300 or 600 extant tips and randomly pruned taxa to leave a tree of 100 or 300 tips, modelling partial sampling over time.

As some scenarios can be distinguished simply by comparing the branch lengths of the corresponding trees, we normalized the time scales so that each of our trees has a mean branch length of 1. This ensures that any differences we observe between the summary statistics in different classes are not simply due to scaling. We did not, however, modify the variances of the branch length distribution, as those may contain some of the signal picked up by summary statistics.

In total, there are 5 scenarios in which we compare trees: HIV/Dengue/Measles (HDM), influenza (2-year USA, 5-year global, 12-year global), HIV contexts (labeled WNH after the first author names of the corresponding publications), simulated trees with 100 tips ('Simulated') and simulated trees with 300 tips ('Simulated300'). Within each set, we performed classification with generalized linear models and random forests, using the tree shape statistics as features. We computed a measure of each feature's importance for each classification. All reconstructed trees, including accession numbers (in the tip labels), have been deposited to `github.com/Leonardini/treeCentralityData`.

3.4 Shape features

We computed a range of topological and spectral summary features of the viral phylogenies (see Table 1 for the definitions and references for each one). Our focus here is on some of the novel tree topology summary statistics, but we also include a number of standard statistics for comparison. All input trees were binary and rooted, and all branch lengths were non-negative, although many of the trees had zero-length branches. All comparisons involved trees on the same number of tips.

For the node properties derived from network science, we focus our discussion on the maximum value of each type of centrality a node can have within a tree, but using other derived statistics (the minimum, mean, median or variance) could also have been an option. For the spectral properties, which also produce a value for each node, we focus on the maximum value as well as the minimum strictly positive value. Lastly, for the distance Laplacian spectral properties we use the four derived statistics proposed by Lewitus and Morlon [36].

Name	Description	Short form	Ref.
Numbers of small configurations			
Cherry number	# of nodes with 2 tip children	cherries	[44]
Pitchforks	# of nodes with 3 tip descendants	pitchforks	[60]
Double cherries	# of nodes with 2 cherry children	doubcherries	new
4-caterpillar	# of caterpillars with 4 tips	fourprong	[60]
Clades of size x	# of nodes with x tip descendants	num$x$	[60]
Tree-wide summaries			
Colless imbalance	Colless imbalance	colless	[12]
Sackin imbalance	Mean path length from tip to root	sackin	[61]
Maximum height	Max # of steps from the root	maxheight	[10]
Maximum width	Max # of nodes at the same depth	maxwidth	[10]
Stairs	Proportion of imbalanced subtrees	stairs	[52]
Max difference in widths	$\max_i(n_{i+1} - n_i)$	delW	[10]
Node properties from network science			
Betweenness centrality	# of shortest paths through node	between	[51]
Weighted betweenness	as above, but with weighted edges	betweenW	[50]
Closeness centrality	total distance to all other nodes	closeness	[51]
Weighted closeness	as above, but with weighted edges	closenessW	[50]
Eigenvector centrality	value in Perron-Frobenius vector	eigen	[51]
Weighted eigenvector	as above, but with weighted edges	eigenW	[50]
Summaries from network science			
Diameter	largest distance between 2 nodes	diameter	[7]
Mean pairwise distance	average distance between 2 nodes	meanpath	[45]
Spectral properties			
Min adjacency	min adjacency matrix eigenvalue > 0	minAdj	[20]
Max adjacency	max adjacency matrix eigenvalue	maxAdj	[20]
Min Laplacian	min Laplacian matrix eigenvalue > 0	minLap	[20]
Max Laplacian	max Laplacian matrix eigenvalue	maxLap	[20]
Distance Laplacian spectral properties			
Max eigenvalue	largest eigenvalue in the spectrum	dLapLambdaMax	[36]
Max density	location of largest spectral density	dLapDensityMax	[36]
Asymmetry	skewness of the spectral density	dLapAsymmetry	[36]
Kurtosis	peakedness of the spectral density	dLapKurtosis	[36]

Table 1: Summary measures for phylogenetic trees. Here, n_i is the number of nodes at depth i.

3.4.1 Tree topology summaries based on network science

Network science, broadly defined as the study of complex networks, has produced a number of tools that have been applied in a variety of contexts [35, 49], including degree sequence, degree assortativity, density, diameter, and a number of node centrality measures. Only some of these apply naturally and informatively to phylogenies. For instance, the degree sequence, the degree centrality, the density and the clustering coefficients do not vary between phylogenetic trees on n tips, while their degree assortativity can take on one of two values which are close to one another for large n (see section 6.4 in the **Supplementary Materials** for details).

We now discuss those network science-inspired features that are informative for phylogenetic trees.

3.4.2 Diameter, average shortest path and Wiener index

The diameter (the maximum length of a shortest path) is a useful summary statistic. For general trees with N nodes, it can be calculated in linear time using a "folklore" dynamic programming algorithm, and its value can vary between 2 for the star and $N-1$ for the path of length N. For phylogenetic trees with $N = 2n-1$ nodes, the range is from $2\log(n)$ to n. The exact maximum and minimum values, as well as the trees attaining them, and the distribution for $N = 45$ ($n = 23$), are described in Figure S10 in **Supplementary Materials**.

The average shortest path length of a tree may also be informative. Unlike for general graphs, it can be calculated in linear time in a tree with N nodes using dynamic programming [45]. The sum of all shortest path lengths between pairs of nodes in a tree is known as the tree's Wiener index, and equals $(2n-1)(n-1)$ times the average shortest path length. The Wiener index for general trees with N nodes is always contained between $(N-1)^2$ and $\binom{N+1}{3}$, values which are attained by the star and the path, respectively [16]. For phylogenetic trees, the range of the Wiener index is substantially narrower; its distribution for $N = 45$ ($n = 23$) is described in Figure S11 in the **Supplementary Materials**. Note that both the diameter and Wiener index generalize naturally to trees with arbitrary positive branch lengths.

Based on their distributions for small phylogenetic trees, shown in Figure S12 in the **Supplementary Materials** we can think of trees that have a diameter or a Wiener index much smaller than the mean as being similar to complete binary trees, and of those that have a diameter or a Wiener index much larger than the mean as being similar to caterpillars or double caterpillars, respectively. This is not too different from the corresponding situation for imbalance statistics such as Colless or Sackin index. However, the situation changes dramatically when we look at node centrality measures defined in network science. A number of such centrality measures exist that can be defined for each node of a graph. In particular, betweenness centrality, closeness centrality and eigenvector centrality are quantities derived from network science that can be computed in linear time for a tree with N nodes and can capture aspects of tree topology not captured by mere asymmetry.

3.4.3 Betweenness centrality

Betweenness centrality associates to each node v in a graph the number of pairs $u, w \in V - \{v\}$ such that the shortest $u - w$ path passes through v; in other words,

$$C_B(v) := \big|\{(u,w) \in V - \{v\} \mid d(u,w) = d(u,v) + d(v,w)\}\big|.$$

Betweenness centrality can be normalized by the number $\binom{N-1}{2}$ of all pairs $u, w \in V - \{v\}$, but we choose not to use this normalization here. In a tree T, there is a unique shortest path between every pair of nodes, and the shortest $u - w$ path passes through v if and only if u and w are in clades subtended by different children of v when T is rerooted at v. Hence, the

betweenness centrality of an internal node v is simply $\Pi_{i<j} n_i n_j$, where $n_1, \ldots, n_k$ are the sizes of the clades subtended by the k children of v when T is rerooted at v, and k is the degree of v in T. This implies that the betweenness centrality of all the nodes in a tree can be computed in linear time, a result that, while not surprising, does not seem to have been mentioned in the literature until now, although an algorithm to perform this computation in a distributed fashion has been published [68]. This contrasts with the situation for a general graph, where the best algorithm takes $O(NM)$ time, where M is the number of edges, making it quadratic time for trees [8].

In a phylogenetic tree, the degree is 1 for a tip, 2 for the root and 3 for internal nodes, so the betweenness centrality is respectively $0, n_1 n_2$ or $n_0 n_1 + n_0 n_2 + n_1 n_2$ in those cases, where n_1 and n_2 are the sizes of the left and right subtrees of the node and n_0 is the number of nodes outside the subtree rooted at this node. This can easily seen to be maximal when $n_0 = n_1 = n_2$, a situation that does not occur in every tree, but is only possible in those with $N \equiv n \equiv 1 \mod 3$. The betweenness centralities are shown for all the nodes of an example tree in Figure 1 (b). Trees with high maximum betweenness centrality are those which have a node whose left subtree, right subtree, and "outside subtree" (what remains of the tree after removing the subtree rooted at the node), are all close in size - a kind of three-way symmetry, as opposed to the two-way (left-right) symmetry measured by classical statistics. The distribution of this quantity, maximum betweenness centrality, for $N = 45$ ($n = 23$), is described in Figure S13 in the **Supplementary Materials**. We also note that the betweenness centrality values of a tree do not change if the tree has variable branch lengths, since each pair of nodes is connected by a unique shortest path. However, this observation is not true for more complicated graphs.

3.4.4 Closeness centrality

Closeness centrality associates to each node v in a graph the inverse of the sum of its distances to all the other nodes in the graph. In other words,

$$C_C(v) := \frac{1}{\sum_u d(u,v)}.$$

The definition means that closeness centrality is inversely proportional to *farness*, the sum of distances from a node to all the other nodes in the graph; hence, it will be small for centrally located nodes and large for remote ones. While this quantity generally requires at least $O(NM)$ time to be computed for a graph with N nodes and M edges [5], this can be reduced to linear time for a tree, an observation that does not seem to have been published previously, although a distributed algorithm for performing this computation has been proposed [69].

Indeed, if we consider an internal node u with a child v, we have $d(v,x) = d(u,x) - 1$ for every node x in the clade of T that v subtends, and $d(v,y) = d(u,y) + 1$ for every node y outside this clade. Hence, by computing the height (distance to the root) of each node and the sizes of the left and right subtrees of each node in a bottom-up traversal of the tree, we can also find the closeness centrality in linear time for all the nodes in the tree. This is illustrated on our example tree in Figure 1 (c). It can also be shown - as we do in Section 6.7 in the **Supplementary Materials** - that the largest value of closeness centrality in a phylogenetic tree is achieved by its unique centroid (the node v such that, if the tree is rerooted at it, no child of v subtends a clade with more than half of all the nodes), while the smallest value is achieved by one of the tips (in fact, this result also holds for trees with arbitrary positive branch lengths). The distribution of this quantity, maximum closeness centrality, for $N = 45$ ($n = 23$), is described in Figure S14 in the **Supplementary Materials**.

3.4.5 Eigenvector centrality

Eigenvector centrality associates to each node v in a connected weighted graph the positive value $e(v)$ such that

$$\sum_{u \sim v} w(uv)e(u) = \lambda e(v)$$

holds for all nodes simultaneously with the largest possible λ. In other words, $\vec{e}$ is the Perron-Frobenius eigenvector of the graph's adjacency matrix. Since this vector is only defined up to a constant, we use the default normalization, which takes it to be a unit vector in the Euclidean norm. We show an alternative normalization, which makes the minimum entry 1, in Figure 1 (d).

Eigenvector centrality is closely related to the adjacency matrix-based spectral methods we discuss below. Since the adjacency matrix of a phylogenetic tree with n tips has exactly $4n - 4$ non-zero entries, it follows that eigenvector centrality can also be computed to a fixed precision ϵ in linear time, with a constant proportional to $\log(1/\epsilon)$. It can also be shown - as we do in Section 6.7 in the **Supplementary Materials** - that the largest value of eigenvector centrality in a phylogenetic tree is achieved by an internal node, though it will not necessarily be the root (this result holds for trees with arbitrary positive branch lengths). The distribution of this quantity, maximum eigenvector centrality, for unweighted (unit branch length) phylogenetic trees with $N = 45$ ($n = 23$), is shown in Figure S15 in the **Supplementary Materials**.

3.4.6 Spectral properties of trees

Spectral properties are properties of the eigenvalues of matrices associated with graphs, such as the adjacency, Laplacian or normalized Laplacian matrices [20]. Since spectral methods are a key tool of algebraic graph theory, it is natural to ask whether they can capture important features of pathogen phylogenies. Some of the eigenvalues are interpretable in a limited number of cases [20]. The eigenvalues of the Laplacian of a graph reveal its number of components, the number of spanning trees it has, and the minimal energy of its balanced orthogonal representation in $\mathbb{R}^m$. The second smallest eigenvalue provides a bound on the *conductance* of the graph, the minimum ratio of the number of edges crossing a cut to the size of the cut. The eigenvalues of the normalized Laplacian also provide information on the conductance of a graph as well as its diameter, number of bipartite components and properties of random walks defined on it [9].

Since trees are bipartite graphs, the spectra of their adjacency matrices are symmetric around 0 [20] and the spectra of their normalized Laplacians are symmetric around 1 [9]. The full set of eigenvalues of a tree's adjacency matrix reveals the number of *matchings* (node-disjoint collections of edges); namely, the characteristic polynomial, which has these eigenvalues as its roots, has coefficients whose absolute values equal the number of matchings of each size [42].

3.4.7 Spectral properties derived from the distance Laplacian matrix

In a recent paper, Lewitus and Morlon [36] used the spectra of the distance Laplacian matrix, obtained by subtracting the distance matrix from a diagonal matrix formed by its row sums, to characterize and compare trees. They found promising results and tentatively stated that these spectra were likely to distinguish trees from one another. In Section 6.8 of our **Supplementary Materials**, on the negative side, we exhibit a pair of trees on $n = 3$ tips with different branch lengths that nevertheless have identical distance Laplacian spectra. On the positive side, we verified that no pair of different phylogenetic trees with at most 32 tips and unit branch lengths has this property; however, we suspect that, as has been shown for most other spectra [42], the fraction of trees on n tips uniquely defined by their distance Laplacian spectra tends to 0 for large n.

We use the four summary statistics proposed by the authors [36] - namely, the maximum eigenvalue, and the asymmetry, kurtosis, and maximum density of the eigenvalue distribution (obtained via smoothing with a Gaussian kernel), as implemented in the *RPANDA* [48] package in R [59]. We note that, unlike the rest of the statistics in this paper, these ones require a number of operations proportional to n^3, where n is the number of tips, and thus take substantially longer to compute.

4 Results

4.1 Tree topology differentiates viruses and epidemiological scenarios

We find that tree topology carries considerable information and differs both between viruses and between different epidemiological scenarios for the same virus, on real as well as simulated data. While degree sequence, clustering coefficients and other measures based in network science are not informative for binary trees, a number of non-standard topological features differ. Furthermore, there are several features that distinguish well between groups of trees with the same overall level of asymmetry, highlighting the need to move beyond asymmetry when using trees to infer evolutionary and epidemiological parameters or to test hypotheses [23]. Figures 2 and 3 illustrate the distributions of all tree statistics considered in this paper. In addition, we perform a (two-sided) Mann-Whitney U test, also known as a Wilcoxon rank-sum test [39], between each pair of scenarios or viruses in a given group, for every statistic we consider, and report the p-values in Tables S2 and S3 in the **Supplementary Materials**. We then apply a Bonferroni correction [15] to account for the multiple hypotheses being tested, and highlight in bold those values which remain significant at the $\alpha = 0.05$ level after this correction.

Most of the statistics do not vary much between the three viruses in Figure 2 (red boxplots). Distinguishing the topologies in these groups of trees requires tools going beyond the traditional symmetry or imbalance metrics; in this case, the only statistically significant difference are produced by the number of cherries, weighted closeness centrality, maximum height, and the proportion of imbalanced subtrees (stairs) capture differences that are not apparent in the imbalance. In contrast, while two of the spectral statistics (maximum adjacency and maximum Laplacian eigenvalues) show statistically significant differences among the groups, the vertical scale is very small, and these differences are a small percentage of the overall value (2% for the maximum adjacency and 1% for the maximum Laplacian).Furthermore, three of the network statistics - namely, the Wiener index (meanpath) and the weighted closeness and eigenvector centralities - also distinguish the three viruses, as do three of the four statistics based on the distance Laplacian.

The green boxplots in Figure 2 show the tree summaries for influenza in three scenarios. It is well-known that global influenza patterns give rise to highly imbalanced trees, an observation which in part motivated the growing field of phylodynamics [23]. In our data, the global five-year and (2-year) USA flu trees have similar, lower, levels of asymmetry. The only statistics which are able to differentiate all three groups at a statistically significant level after multiple testing correction are betweenness centrality and the maximum Laplacian eigenvalue. As with the three viruses, the differences in the latter are not pronounced, and are very small in relative magnitude. In addition, the LM spectral statistics are quite discriminating for these scenarios, though none of them is able to do so in a statistically significant way. Figure 6 shows a random tree from each group, and differences between the shapes of the flu trees are not immediately apparent by eye.

The blue boxplots in Figure 2 show the summaries for small trees (100 tips) in the four simulated settings. This time, only the closeness centrality, the diameter, and the mean path, as well as the number of cherries, the Sackin and Colless imbalances, the maximum height, the

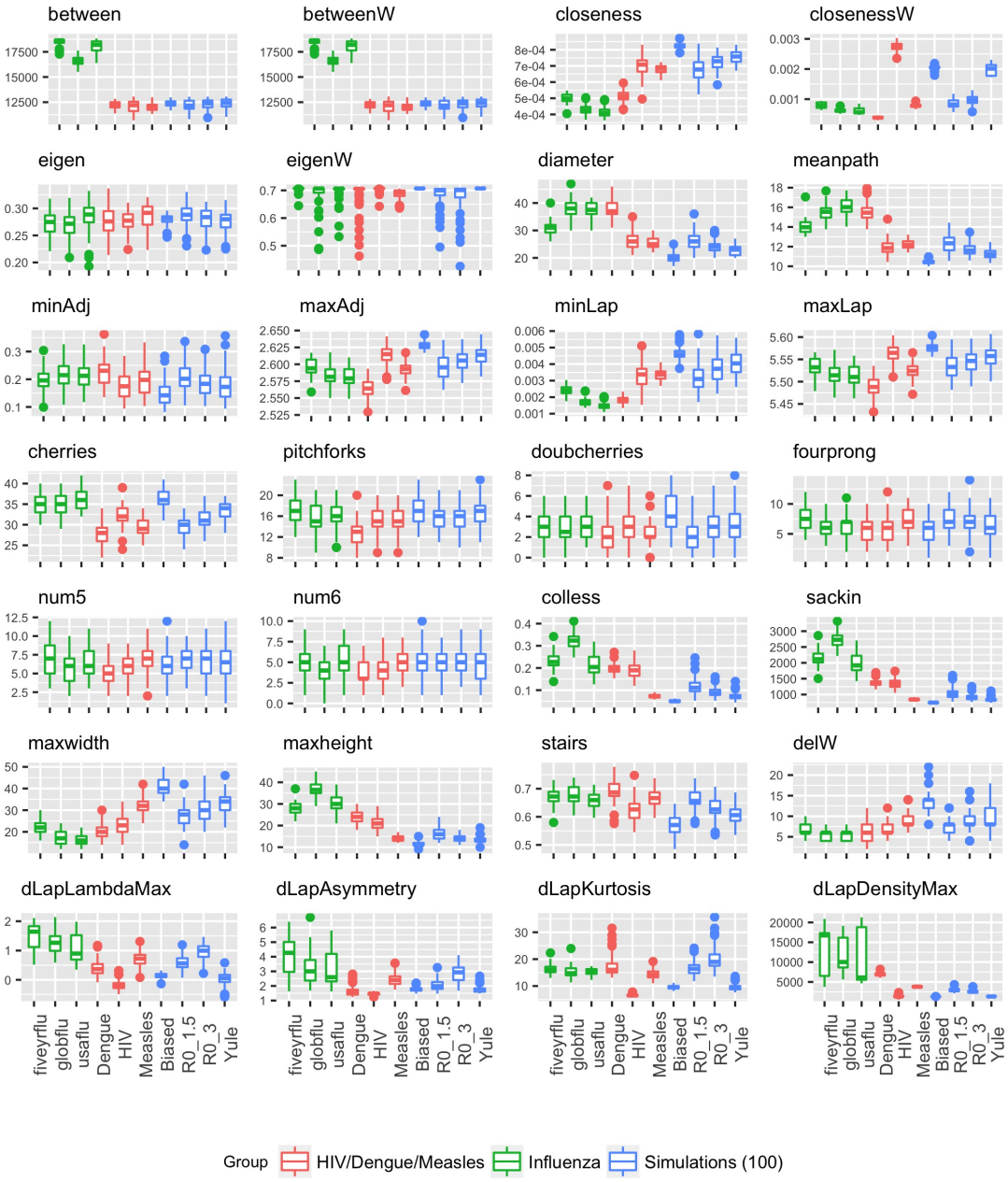


Figure 2: Tree summaries for the HIV/Dengue/Measles, influenza and simulations of size 100

stairs statistic and the lambdaMax statistic based on the distance Laplacian spectrum, are able to differentiate every pair of these scenarios in a statistically significant way.

In contrast, the cyan boxplots in Figure 3 show that large trees in the same simulated settings can be discriminated by all of these statistics, but also by several other ones, including the adjacency and Laplacian eigenvalues as well as the maximum width and the maximum difference

Figure 3: Tree summaries for HIV in three settings and simulations of size 300

in width (delW). This suggests that it is strictly easier to discriminate larger trees than it is to discriminate smaller trees.

Lastly, the three epidemiological scenarios (concentrated epidemic, generalized epidemic in a village, and generalized epidemic in a country) for HIV, shown in the red boxplots in Figure 3, appear to be quite difficult to distinguish. Only the weighted and unweighted closeness, as

well as the Sackin and Colless imbalance, the maximum eigenvalues of the adjacency and the Laplacian, and the asymmetry and kurtosis of the distance Laplacian spectrum, are able to do so in a statistically significant manner.

Figure 4 shows the mutual information between the virus or epidemiological scenario and the value of each statistic, for the 5 settings we tested. We have computed these values by discretizing the range of each tree shape statistic into 20 equally-sized bins. The first panel shows that the highest mutual information with the virus (HIV, Dengue or Measles) is obtained by a network science-based statistic in Figure 4, namely, the weighted closeness, with the densityMax coming close second, while the second panel shows that the highest mutual information with the type of flu (global, US or 5-year) is obtained by the minimum Laplacian eigenvalue, with the betweenness (which is the same whether it is weighted or unweighted) and densityMax coming close behind. In the remaining three scenarios (third, fourth and fifth panels), closeness centrality is the only statistic consistently in the top 3 highest mutual information with the model category or scenario.

4.2 The addition of network statistics improves classification

In some scenarios, the trees are markedly different and only a few features are necessary to distinguish them. For example, in the influenza scenario it is clear from Figure 3 that the Colless, Sackin, betweenness and maximum heights statistics all distinguish between global influenza and the other two groups, whereas the closeness measures, the diameter, the mean path and the maximum width distinguish the five year trees from the other two groups. Asymmetry measures alone do not distinguish the three groups.

We use a generalised linear model and two flavours of random forests to classify trees within each scenario. For example, we classified trees as HIV, Dengue or Measles; we classified influenza trees as five-year, global, and USA; we classified simulated trees as biased, Yule, $R_0 = 1.5$ or 3; and we classified the HIV trees by epidemiological scenario. In each classification task, we randomly selected 75% of the trees (75 trees from each group) for training and used the remaining 25 trees from each group for testing, and computed the classification error of the predictor on the testing set. We report the overall classification error with and without the features based on network science, as well as with and without the features based on the distance Laplacian spectrum (abbreviated as "LM" statistics after the authors Lewitus and Morlon). The results are shown in Figure 5 (a). It is clear that the standard ("basic") tree shape statistics are not able to get close to perfect classification on any of the datasets except for the (relatively simple) task of distinguishing three different viruses. The addition of the costly LM statistics improves the performance, but so does the addition of the easily computable new network science-based statistics we introduce in this paper, with comparable gains in performance relative to the baseline (a larger improvement on the flu scenarios, and a slightly smaller one on the HIV scenarios and on simulated data). Interestingly, the addition of network statistics actually increases the error relative to the baseline on the large simulated trees, an effect that is likely due to the random nature of the classifier.

Both random forests and the GLM regression additionally provide estimates of the importance of each feature. For the random forest classifier, the prediction accuracy on the out-of-bag portion of the data is recorded for each tree, and then the same is done after randomly permuting each predictor variable. The differences between the two accuracies are then averaged over all trees, and normalized by the standard error. For the GLM regression, the absolute value of the t-statistic for each model parameter is used as the importance measure. We used the *caret* package [34] in R [59] to compute these for all the classifiers. For each classification task in each scenario, we then ranked the features by importance. We show the rank data in Figure 5 (b). It is apparent that weighted closeness centrality is the feature with the highest importance (lowest rank) across all scenarios, on average, with the much costlier to compute distance Laplacian-based features coming close behind.

5 Conclusions

We have used network science to developed tailored statistics to capture the topologies of phylogenetic trees and illustrated how they compare across different viruses and scenarios. The viruses vary in pathogenesis and their rates and modes of transmission, characteristics which ideally should be reflected in the tree shape. We find that network-science-derived statistics are complementary to asymmetry and frequency of small configurations with a tree. We have also used a tree kernel to compare trees; while it distinguishes well between three different viruses, the statistics do a better job of distinguishing the three influenza scenarios. However, no individual statistic is sufficient to consistently capture the diversity of tree topologies.

It is highly likely that investigators will need to use multiple summary measures to resolve biologically meaningful differences among phylogenies. This is a problem of *feature selection*. Indeed, some of the literature on phylogenetic tree topologies has focused on evaluating which summary measures are the best suited for detecting informative differences in tree topologies [31, 47, 1]. This effort makes the implicit assumption that there exists a specific subset of features that will always be a good choice for all data sets (or data for a given domain), which is not necessarily the case. There may exist summary measures that have not been formulated but would confer an optimal fit to a given dataset. Matsen [41] proposed an innovative approach to this problem, where a generalized and flexible class of tree topology statistics (which can express many established measures of tree shape) can be adapted to the data using a genetic algorithm.

The underlying evolutionary or epidemiological process may affect tree topologies, but they are likely also to be affected by sampling density and uniformity, geographical constraints and population structure, ecological dynamics and host contact structure (in the case of pathogens). Our results point to the important role of tree topologies as a vehicle for obtaining information about some of these underlying processes. Topological summary features are readily interpretable, easy to compute and scale readily to large trees, even when they do not necessarily provide direct insights into the underlying evolutionary or diversification processes. The network-based and spectral statistics we have described also generalize naturally to phylogenetic networks.

On the other hand, each feature can capture only a narrowly-defined aspect of tree shape, such as imbalance. This could potentially be overcome by using multiple features, although correlations among features may lead to diminishing returns with the addition of new features. Features can also be difficult to compare among trees of different sizes (numbers of tips), and despite attempts to derive normalized statistics they remain sensitive to size. However, Pompei and colleagues [55] have proposed a randomization procedure to derive the null distribution of summary features as a possible resolution of this problem.

Acknowledgments: This work was in part supported by the AWS Cloud Credits for Research program. The authors would like to thank Nithum Thain and Stepan Grinek for providing computational support to this project, and Kamyar Khodamoradi for pointing out the theorem by Jordan relevant to the existence of centroids. LC was supported by an NSERC Discovery Grant (RGPIN-2016-04622) as well as an Alfred P. Sloan Fellowship (FG-2016-6392). LC also acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. CC was supported by the Engineering and Physical Sciences of the United Kingdom (EP/K026003/1 and EP/N014529/1). AP was supported by an NSERC Discovery Grant (RGPIN-2018-05516) as well as a CIHR Project Grant (PJT-155990). This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program (CC).

References

[1] P.-M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Syst Biol*, 51(6):866–72, Dec 2002.

[2] D. Aldous. Probability distributions on cladograms. In *Random Discrete Structures*, The IMA Volumes in Mathematics and its Applications, pages 1–18. Springer New York, 1996.

[3] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16(1):23–34, 1 Feb. 2001.

[4] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 33:D34–D38, 2005.

[5] E. Bergamini, M. Borassi, P. Crescenzi, A. Marino, and H. Meyerhenke. *Computing Top-k Closeness Centrality Faster in Unweighted Graphs*, pages 68–80. SIAM, 2016.

[6] M. G. Blum and O. François. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006.

[7] B. Bollobás. *Modern Graph Theory*. Springer, 2013.

[8] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[9] F. R. K. Chung. *Spectral Graph Theory*. University of Pennsylvania - AMS, 1997.

[10] C. Colijn and J. Gardy. Phylogenetic tree shapes resolve disease transmission patterns. *Evol Med Public Health*, 2014(1):96–108, 9 June 2014.

[11] C. Colijn and G. Plazzotta. A metric on phylogenetic tree shapes. *Systematic Biology*, 67:113–126, 2018.

[12] D. H. Colless. Relative symmetry of cladograms and phenograms: an experimental study. *Syst. Biol.*, 1995.

[13] M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, pages 625–632, 2001.

[14] K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410 – 418, 2010.

[15] O. J. Dunn. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, 30(1):192–197, Mar. 1959.

[16] R. C. Entringer, D. E. Jackson, and D. A. Snyder. Distance in graphs. *Czechoslovak Math. J.*, 26:283–296, 1976.

[17] T. Flouri, F. Izquierdo-Carrasco, D. Darriba, A. Aberer, L.-T. Nguyen, B. Minh, A. Von Haeseler, and A. Stamatakis. The phylogenetic likelihood library. *Systematic Biology*, 64(2):356, 2015.

[18] B. Foley, T. Leitner, C. Apetrei, B. Hahn, I. Mizrachi, J. Mullins, A. Rambaut, S. Wolinsky, and B. Korber. HIV Sequence Compendium 2013. Technical Report LA-UR 13-26007, Los Alamos National Laboratory, NM, 2013.

[19] G. Fusco and Q. Cronk. A new method for evaluating the shape of large phylogenies. *J Theor Biol*, 175(2):235–243, 21 July 1995.

[20] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer, 2001.

[21] A. J. Goldman. Optimal center location in simple networks. *Transportation Science*, 5(2):212–221, 1971.

[22] P. A. Goloboff, J. S. Arias, and C. A. Szumik. Comparing tree shapes: beyond symmetry. *Zoologica Scripta*, 46(5):637–648, 2017.

[23] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–32, Jan 2004.

[24] C. Guyer and J. Slowinski. Adaptive radiation and the topology of large phylogenies. *Evolution*, 47(1):253–263, 1993.

[25] M. D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *Journal of Classification*, 10(1):5–24, 1993.

[26] D. M. Hillis, T. A. Heath, and K. S. John. Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482, 2005.

[27] K. T. Huber, A. Spillner, R. Suchecki, and V. Moulton. Metrics on multilabeled trees: Interrelationships and diameter bounds. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 8(4):1029–1040, Jul. 2011.

[28] G. Hunt, J. Ledwaba, A. Salimo, M. Kalimashe, B. Singh, A. Puren, and L. Morris. Surveillance of transmitted HIV-1 drug resistance in 5 provinces in South Africa in 2011. *Communicable Diseases Surveillance Bulletin*, 11:122–124, 2013.

[29] C. Jordan. Sur les assemblages de lignes. *Journal für reine und angewandte Mathematik*, 70:185–190, 1869.

[30] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

[31] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, pages 1171–1181, 1993.

[32] S. L. Kosakovsky Pond, D. Posada, E. Stawiski, C. Chappey, A. F. Poon, G. Hughes, E. Fearnhill, M. B. Gravenor, A. J. Leigh Brown, and S. D. Frost. An Evolutionary Model-Based Algorithm for Accurate Phylogenetic Breakpoint Mapping and Subtype Prediction in HIV-1. *PLOS Computational Biology*, 5(11):1–21, 11 2009.

[33] M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[34] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-78.

[35] T. G. Lewis. *Network Science: Theory and Applications*. Wiley, 2009.

[36] E. Lewitus and H. Morlon. Characterizing and comparing phylogenies from their Laplacian spectrum. *Systematic Biology*, 65(3):495–507, 2016.

[37] G. Li. Generation of rooted trees and free trees. Master's thesis, University of Victoria, 1996.

[38] M. Manceau, A. Lambert, and H. Morlon. Phylogenies support out-of-equilibrium models of biodiversity. *Ecol Lett*, 18(4):347–356, 2015.

[39] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar. 1947.

[40] F. Matsen. A geometric approach to tree shape statistics. *Syst Biol*, 55(4):652–661, Aug. 2006.

[41] F. Matsen. Optimization over a class of tree shape statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(3):506–512, 2007.

[42] F. A. Matsen and S. N. Evans. Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials. *Algorithms for Molecular Biology*, 14(7), 2012.

[43] B. D. McKay. On the spectral characterisation of trees. *Ars Combinatoria*, 3:219–232, 1977.

[44] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164(1):81–92, 2000.

[45] B. Mohar and T. Pisanski. How to compute the Wiener index of a graph. *Journal of Mathematical Chemistry*, 2:267–277, 1988.

[46] M. B. Monagan, K. O. Geddes, K. M. Heal, G. Labahn, S. M. Vorkoetter, J. McCarron, and P. DeMarco. *Maple 10 Programming Guide*. Maplesoft, Waterloo ON, Canada, 2005.

[47] A. Mooers and S. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, 1997.

[48] H. Morlon, F. Condamine, E. Lewitus, and M. Manceau. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. *Methods in Ecology and Evolution*, 2016. R package version 1.0.

[49] M. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. The Princeton Press, 2006.

[50] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70:056131, Nov. 2004.

[51] M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[52] M. M. Norström, M. C. F. Prosperi, R. R. Gray, A. C. Karlsson, and M. Salemi. PhyloTempo: A set of R scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. *Evol. Bioinform. Online*, 8:261–269, 11 June 2012.

[53] V. Novitsky, H. Bussmann, A. Logan, S. Moyo, E. van Widenfelt, L. Okui, M. Mmalane, J. Baca, L. Buck, E. Phillips, and D. Tim. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. *PLoS One*, 8:e80589, 2013.

[54] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[55] S. Pompei, V. Loreto, and F. Tria. Phylogenetic properties of RNA viruses. *PLOS One*, 7(9):e44849, 20 Sept. 2012.

[56] A. F. Poon. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Mol Biol Evol*, 32(9):2483–2495, Sept. 2015.

[57] A. F. Poon, L. W. Walker, H. Murray, R. M. McCloskey, P. R. Harrigan, and R. H. Liang. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One*, 8(11):e78122, 2013.

[58] A. Purvis, S. A. Fritz, J. Rodríguez, P. H. Harvey, and R. Grenyer. The shape of mammalian phylogeny: patterns, processes and scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1577):2462–2477, 2011.

[59] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[60] N. A. Rosenberg. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. *Annals of Combinatorics*, 10(1):129–146, 2006.

[61] M. Sackin. "Good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

[62] E. Saulnier, O. Gascuel, and S. Alizon. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS Comput. Biol.*, 13(3):e1005416, 6 Mar. 2017.

[63] J. Slowinski. Probabilities of n-trees under two models: A demonstration that asymmetrical interior nodes are not improbable. *Syst Zool*, 39(1):89–94, 1990.

[64] E. Stam. Does imbalance in phylogenies reflect only bias? *Evolution*, 56(6):1292–1295, June 2002.

[65] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[66] M. Stich and S. Manrubia. Topological properties of phylogenetic trees in evolutionary models. *Eur Phys J B*, 70(4):583–592, 21 July 2009.

[67] J. Trifinopoulos, L.-T. Nguyen, A. von Haeseler, and B. Minh. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44:W232–W235, 2016.

[68] W. Wang and C. Y. Tang. Distributed computation of node and edge betweenness on tree graphs. In *52nd IEEE Conference on Decision and Control*, pages 43–48, Dec. 2013.

[69] W. Wang and C. Y. Tang. Distributed computation of classic and exponential closeness on tree graphs. In *Proceedings of the American Control Conference*, Jun. 2014.

[70] E. Wolf, J. Herbeck, S. Van Rompaey, M. Kitahata, K. Thomas, G. Pepper, and L. Frenkel. Phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. *AIDS Research and Human Retroviruses*, 33:318–22, 2017.

[71] T. Wu and K. Choi. On joint subtree distributions under two evolutionary models. *Theor Popul Biol*, 108:13–23, 29 Nov. 2015.

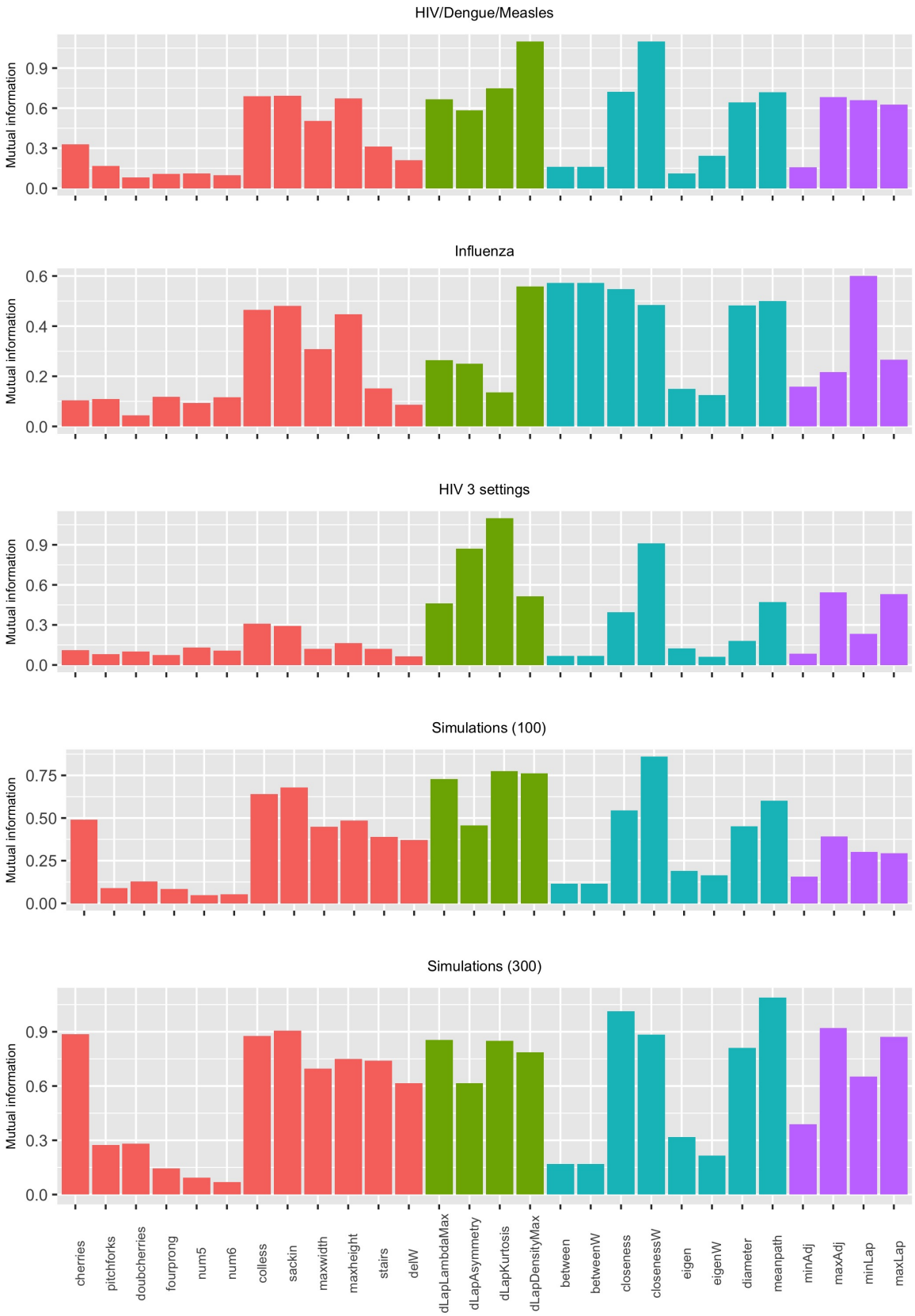
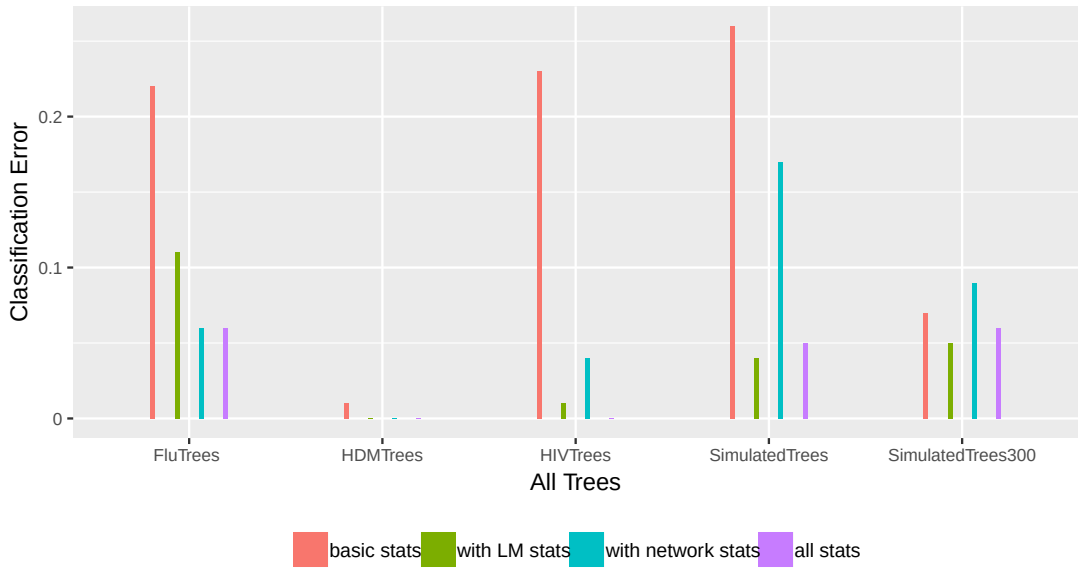
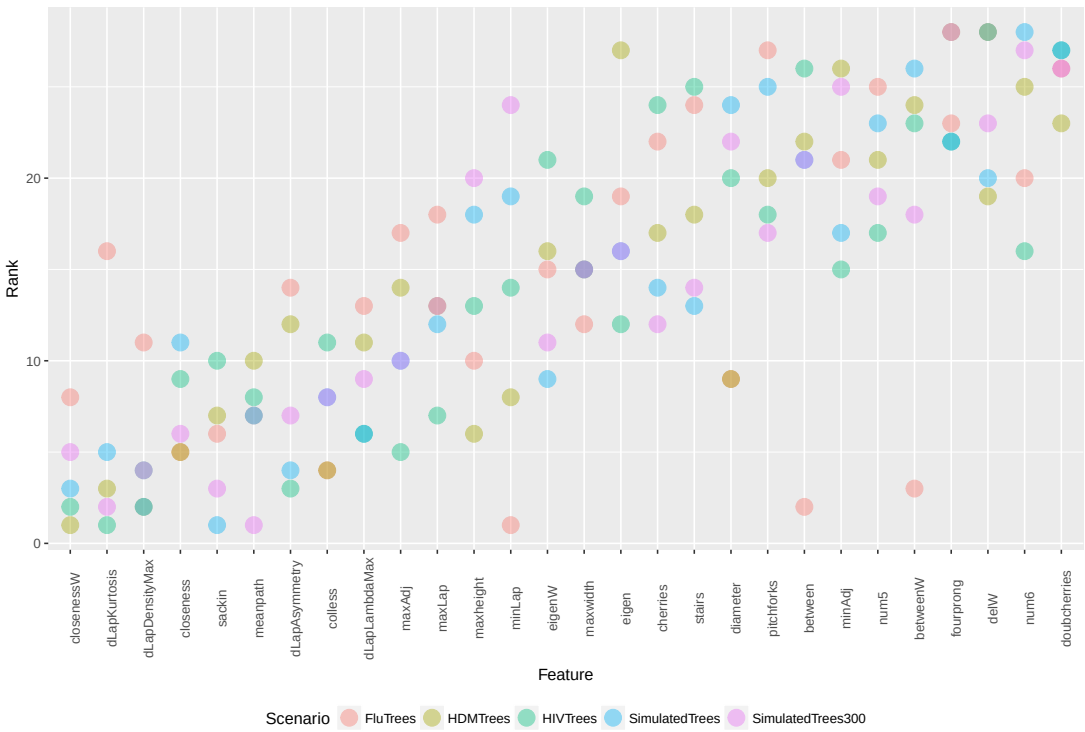


Figure 4: Mutual information between tree summaries and the virus type or epidemiological scenario. Labels are as in Table 1, and the colour indicates the type of statistic as per Table 1, i.e. red: basic statistics; green: distance Laplacian spectrum statistics; blue: network science-based statistics; purple: spectral statistics.



(a)



(b)

Figure 5: (a) Accuracy of classification with and without network science features, as well as with and without the distance Laplacian spectral features. (b) Feature importance in multi-class classification across all scenarios, ordered by the median. Here the importance for each feature is computed based on the mean decrease in node impurity (we use *varImp* function [33] in *R* which uses Gini index as an an impurity function)
. Each point is the rank of the corresponding feature in one of the classification tasks. Low ranks correspond to the most important features (i.e. the top-ranked feature has rank 1, and so on).

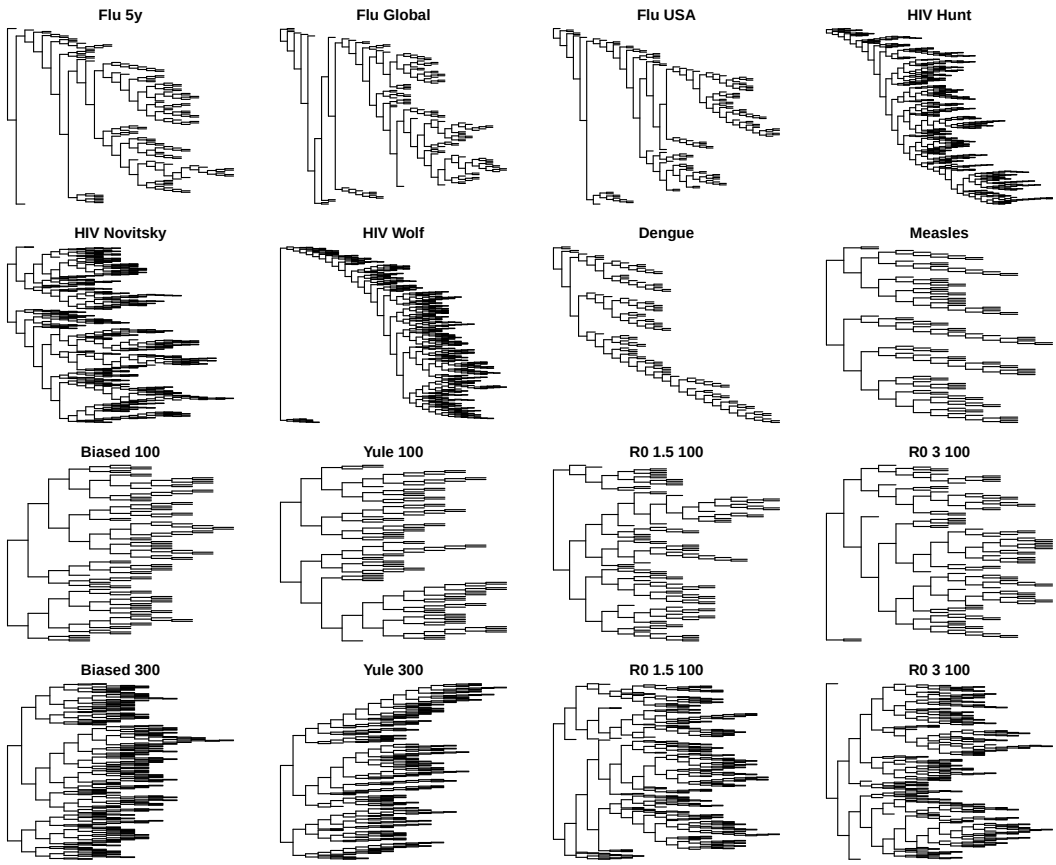


Figure 6: A randomly sampled tree from each scenario (except HIV in the 3-virus comparison because HIV is represented in three other trees). To allow for focus on tree shape rather than on branch lengths, trees have been visualized with branch lengths set to 1.

6 Supplementary Materials

6.1 GenBank accession numbers for outgroups

Virus	Outgroup	GenBank accession number
HIV-1 subtype B	a subtype D sequence	AY071949
Dengue virus serotype 4	isolate from the Philippines (1956)	U18433
Measles virus	a genotype D6 sequence	AY523581

Table S1: Accession numbers for outgroups

6.2 p-values for the statistics used to compare viruses or scenarios

The p-values come from the Mann-Whitney U test for each statistic in a pairwise comparison; the values passing Bonferroni correction for multiple testing at the $\alpha = 0.05$ level are in bold.

	fiveyrflu-globflu	globflu-usaflu	fiveyrflu-usaflu	Dengue-HIV	Dengue-Measles	HIV-Measles	Yule-Biased	Yule-R0_1.5	Yule-R0_3	Biased-R0_1.5	Biased-R0_3	R0_1.5-R0_3
between	**1.87e-38**	**9.49e-23**	**1.73e-05**	0.116	**3.74e-05**	0.137	0.724	0.0114	0.338	0.00809	0.427	0.097
betweenW	**1.87e-38**	**9.49e-23**	**1.73e-05**	0.116	**3.74e-05**	0.137	0.724	0.0114	0.338	0.00809	0.427	0.097
closeness	**1.03e-22**	0.0357	**3.29e-25**	**1.49e-62**	**8.12e-96**	0.00267	**1.01e-30**	**1.27e-19**	**1.74e-07**	**1.09e-41**	**1.91e-40**	**5.23e-08**
closenessW	**5.71e-17**	0.546	**4.86e-14**	**1.4e-136**	**5.85e-118**	**1.91e-142**	0.001	**8.22e-109**	**5.07e-100**	**5.04e-107**	**1.83e-95**	**1.26e-07**
eigen	0.399	0.012	0.0526	0.934	0.00477	0.00112	0.0658	**2.72e-05**	0.105	0.000363	0.795	0.00789
eigenW	0.00523	0.163	0.0573	0.00169	0.432	**1.61e-18**	0.101	**1.51e-08**	**1.42e-07**	**1.52e-08**	**1.43e-07**	0.353
diameter	**7.79e-21**	0.354	**5.08e-22**	**1.46e-67**	**1.12e-74**	0.0595	**1.13e-23**	**3.42e-17**	**1.63e-06**	**1.89e-39**	**2.48e-35**	**8.73e-07**
meanpath	**2.35e-15**	0.00199	**9.51e-21**	**1.54e-81**	**2.34e-73**	0.000369	**2.03e-36**	**1.35e-21**	**1.81e-09**	**1.14e-40**	**6.41e-42**	**4.58e-09**
minAdj	0.0525	0.608	0.18	**6.87e-12**	**4.44e-06**	0.0222	**5.82e-05**	**5.36e-06**	0.0873	**1.39e-16**	**7.61e-09**	0.00262
maxAdj	**3.07e-05**	0.62	**4.22e-06**	**6.74e-64**	**2.11e-37**	**2.78e-28**	**2.9e-20**	**5.01e-13**	**4.66e-05**	**1.03e-33**	**3.95e-30**	**8.35e-05**
minLap	**6.37e-23**	**6.63e-06**	**2.41e-30**	**2.19e-39**	**3.17e-97**	0.587	**4.04e-12**	**8.29e-13**	0.00143	**7.43e-33**	**2.51e-21**	**1.21e-05**
maxLap	**9.17e-05**	0.855	**3.64e-05**	**1.47e-59**	**3.57e-27**	**1.43e-33**	**1.18e-14**	**3.6e-09**	0.00221	**3.43e-28**	**1.55e-23**	0.000958
cherries	0.515	0.00773	0.0268	**5.79e-28**	**2.35e-05**	**4.45e-20**	**1.15e-19**	**1.25e-31**	**9.33e-11**	**1.1e-55**	**3.02e-36**	**1.23e-10**
pitchforks	0.0122	0.765	0.0283	**7.42e-14**	**1.08e-13**	0.565	0.0491	0.00465	0.0103	**3.37e-06**	**1.03e-05**	0.77
doubcherries	0.458	0.0415	0.192	**1.31e-07**	0.0368	0.000347	0.000208	**2.51e-06**	0.119	**1.33e-13**	**1e-06**	0.00227
fourprong	**0.000101**	0.582	0.00295	0.774	**3.61e-08**	**1.17e-08**	0.0113	0.00175	0.284	**1.21e-07**	0.000595	0.0448
num5	0.0057	0.00411	0.962	3e-04	**2.35e-13**	**6.51e-05**	0.693	0.485	0.387	0.237	0.186	0.821
num6	0.000614	0.000228	0.505	0.000557	**2.06e-13**	**6.3e-05**	0.459	0.239	0.0362	0.662	0.177	0.363
colless	**2.36e-21**	**5.73e-20**	0.125	0.0918	**5.23e-87**	**1.52e-59**	**1.15e-28**	**1.26e-22**	**3.75e-11**	**3.46e-36**	**9.93e-38**	**3.44e-10**
sackin	**3.37e-20**	**2.29e-20**	0.0378	0.469	**4.17e-81**	**2.09e-57**	**6.98e-32**	**3.85e-22**	**7.95e-11**	**3.24e-35**	**3.82e-36**	**2.82e-10**
maxwidth	**1.19e-13**	0.238	**2.53e-17**	**2.88e-05**	**4.95e-64**	**1.65e-34**	**1.06e-29**	**5.96e-15**	**7.32e-05**	**1.01e-58**	**2.39e-42**	**4.04e-05**
maxheight	**1.86e-25**	**1.2e-12**	0.000156	**3.36e-09**	**1.11e-73**	**3.05e-47**	**3.21e-27**	**1.77e-18**	**1.04e-06**	**2.65e-38**	**2.62e-37**	**1.37e-08**
stairs	0.184	0.00291	0.0675	**1.1e-27**	**2.01e-07**	**1.39e-15**	**2.57e-14**	**1.47e-26**	**2.35e-07**	**1.22e-45**	**5.26e-26**	**1.37e-08**
delW	0.000534	1	0.000174	0.124	**5.27e-21**	**2.73e-18**	**9.47e-25**	**5.35e-10**	0.00257	**8.65e-49**	**4e-35**	0.00118
lambdaMax	0.00265	0.0566	**2.79e-05**	**4.35e-48**	**2.58e-16**	**4.44e-69**	**3.85e-07**	**1.39e-42**	**1.55e-68**	**4.74e-36**	**1.97e-56**	**9.91e-23**
asymmetry	0.000291	0.865	0.000994	**5.78e-07**	**1.25e-36**	**1.03e-46**	0.175	**1.96e-11**	**3.15e-36**	**2.38e-10**	**1.45e-34**	**1.03e-24**
kurtosis	0.014	0.891	0.00115	**2.15e-53**	**1.33e-11**	**2.08e-74**	0.91	**2.54e-53**	**3.66e-48**	**2.09e-50**	**9.32e-47**	**1.05e-12**
densityMax	0.152	0.237	0.0281	**1.1e-132**	**3.4e-99**	**2.19e-148**	0.0113	**1.11e-60**	**1.22e-54**	**1.04e-60**	**1.11e-54**	**1.16e-10**

Table S2: p-values for pairs of scenarios in Figure 2.

	Wolf-Hunt	Hunt-Novitsky	Wolf-Novitsky	Yule_300-Biased_300	Yule_300-R0_1.5_300	Yule_300-R0_3_300	Biased_300-R0_1.5_300	Biased_300-R0_3_300	R0_1.5_300-R0_3_300
between	0.613	0.0864	0.203	0.955	0.931	0.454	0.955	0.355	0.491
betweenW	0.613	0.0864	0.203	0.955	0.931	0.454	0.955	0.355	0.491
closeness	**2.68e-29**	**3.38e-05**	**3.24e-40**	**5.58e-51**	**2.93e-59**	**9.48e-38**	**1.15e-70**	**1.63e-66**	**6.47e-23**
closenessW	**1.48e-37**	**4.25e-90**	**3.93e-122**	**1.6e-05**	**1.88e-110**	**2.52e-104**	**1.07e-142**	**2.02e-151**	**1.04e-07**
eigen	0.305	0.774	0.599	0.0138	**1.75e-11**	0.000752	**5.84e-10**	0.026	0.000536
eigenW	0.479	0.0736	0.126	0.612	**7.6e-08**	**5.14e-09**	**7.6e-08**	**5.14e-09**	0.947
diameter	**5.07e-14**	0.919	**8.16e-15**	**1.01e-36**	**4.25e-42**	**3.07e-22**	**2.34e-56**	**6.84e-53**	**3.1e-18**
meanpath	**5.58e-34**	0.000281	**2.42e-45**	**7.26e-56**	**2.38e-50**	**8.81e-41**	**1.28e-59**	**8.72e-61**	**1.04e-25**
minAdj	0.000235	0.369	0.00246	**3.88e-05**	**3.49e-28**	**4.36e-14**	**7.71e-42**	**6.38e-27**	**4.32e-06**
maxAdj	**1.92e-37**	**3.46e-11**	**4.62e-60**	**8.73e-39**	**2.17e-60**	**1.24e-35**	**4.59e-67**	**4.83e-56**	**3.41e-21**
minLap	**5.28e-13**	0.0089	**6.73e-19**	**1.6e-26**	**1.24e-33**	**1.26e-15**	**8.98e-64**	**6.49e-55**	**2.23e-10**
maxLap	**5.29e-35**	**1.06e-10**	**2.35e-57**	**7.6e-32**	**1.13e-55**	**6.74e-30**	**2.21e-63**	**2.75e-51**	**2.67e-19**
cherries	0.0149	0.00883	**3.52e-07**	**3.5e-39**	**8.79e-70**	**4.92e-37**	**3.2e-99**	**1.23e-73**	**2.93e-20**
pitchforks	**8.5e-05**	0.00429	0.226	**0.000119**	**5.88e-15**	**5.75e-07**	**6.3e-32**	**2.22e-20**	0.000248
doubcherries	0.00515	0.0178	**3.87e-07**	**1.56e-08**	**3.4e-14**	0.000252	**2.95e-32**	**2.86e-19**	**1.61e-06**
fourprong	0.00603	0.00224	0.799	**1.41e-07**	0.0269	0.032	**2.67e-12**	**6.59e-12**	0.972
num5	**6.35e-05**	0.324	**2.7e-07**	0.517	0.02	0.641	0.12	0.842	0.0669
num6	0.317	0.00385	**1.35e-05**	0.118	0.416	0.902	0.468	0.0725	0.325
colless	**1.37e-05**	**1.12e-14**	**9.28e-24**	**4.94e-31**	**3.56e-36**	**6.08e-26**	**4.88e-44**	**2.58e-45**	**2.95e-18**
sackin	**4.45e-06**	**8.42e-14**	**2.15e-23**	**4.42e-35**	**4.05e-35**	**3.86e-27**	**4.32e-43**	**9.3e-46**	**6.1e-18**
maxwidth	**4.52e-06**	0.000964	0.0565	**1.35e-47**	**5.52e-31**	**6.92e-08**	**4.03e-93**	**2.82e-68**	**2.42e-15**
maxheight	0.655	**2.3e-13**	**9.51e-13**	**6.52e-32**	**1.19e-37**	**1.25e-16**	**4.68e-51**	**1.54e-45**	**2.8e-18**
stairs	0.000931	0.00214	**7.94e-10**	**8.73e-28**	**5.75e-58**	**1.97e-25**	**4.01e-87**	**5.18e-61**	**2.4e-20**
delW	0.00287	0.119	0.105	**6.46e-47**	**5.19e-22**	**3.22e-07**	**1.86e-81**	**2.14e-65**	**3.27e-08**
lambdaMax	0.122	**2.19e-37**	**2.63e-46**	**6.85e-08**	**2.14e-64**	**6.19e-79**	**1.71e-53**	**5.17e-64**	**1.07e-22**
asymmetry	**1.09e-34**	**2.98e-84**	**6.55e-105**	0.000942	**6.06e-29**	**9.41e-36**	**4.44e-25**	**3.74e-34**	**5.38e-20**
kurtosis	**1.71e-102**	**3.2e-110**	**1.35e-156**	0.0436	**1.14e-58**	**8.03e-59**	**1.58e-55**	**2.88e-57**	**8.21e-25**
densityMax	**4.56e-40**	0.00644	**7.98e-55**	0.0114	**9.21e-72**	**2.61e-72**	**9.09e-72**	**2.56e-72**	**4.45e-15**

Table S3: p-values for pairs of scenarios in Figure 3.

6.3 Kernel-based tree comparisons

Rather than using summary statistics to quantify differences in tree topologies, we can calculate a similarity or distance measure between trees. For example, the Robinson-Foulds distance measure is an edit distance between trees with the same labels (*i.e.*, relating the same taxa). The resulting distance matrix for a set of tree topologies can then be interpreted as a measure space for supervised or unsupervised classifiers [26]. However, the requirement for shared labels in the Robinson-Foulds distance prevents its application to the present problem of comparing phylogenies from different viruses.

To overcome this limitation, we previously adapted a kernel method from natural language processing [13] to provide a similarity measure that operates on both the topology and branch lengths of trees [57]. Every tree is comprised of a large number of subset trees that can each act as a feature. A subset tree is a contiguous set of branches rooted at an internal node of its parent tree, which does not necessarily include all descendants of that internal node. In other words, the subset tree does not have to extend out to the tips of the parent tree.

A subset tree is completely defined by its branching order if we rotate branches of the tree (so-called "ladderization") so that all branching events occur preferentially to one side. For a tree of even modest size, the number of all possible subset trees is extremely large; the space of all possible subset trees is even more immense. Clearly, it is not feasible to exhaustively enumerate the appearance of every possible subset tree for a given observed tree.

The kernel trick is a well-established technique in machine learning that provides an efficient way to compare trees with respect to their subset trees by limiting the comparison to those features that appear in one or both trees [57]. Calculating the inner products over this restricted set of features for every pair of trees yields a distance matrix which defines a projection of the trees into a high dimensional space with convenient properties for machine learning.

We applied this kernel method to the data sets examined above with the following parameter settings: decay factor $\lambda = 0.2$; radial basis function variance $\sigma = 2$.

Figure S1 represents a set of principal component analysis (PCA) plots that illustrate the separation of HIV, Dengue, and Measles virus phylogenies into distinct clusters in the space defined by the tree kernel. The HIV trees are separated from the others along the first principal component that explained roughly 75% of the overall variation. Dengue and Measles virus trees can only be separated into distinct clusters by the third principal component. For comparison, the first two principal components of the tree shape statistics shown in Figure S2 are already able to separate the three viruses. The dots represent individual trees; the arrows point in the direction of maximum multiple correlation with the principal components for each tree shape statistic, while their length indicates the strength of this multiple correlation. Once again, we see that the network science-based statistics are overrepresented among the longest arrows.

Furthermore, applying the same method to trees from the three different samples of influenza A virus outbreaks was less promising (Figure S4). While it is possible that adjusting the kernel tuning parameters (λ and σ) could yield better results, we would risk over-fitting this method to these data. In contrast, Colless and Sackin imbalance separate global flu from the other two groups, while closeness, diameter and mean path separate the five-year flu trees from the global and USA groups. A combination of these summary statistics separate the three groups, as shown in Figure S3 below.

We observed a similar situation in the case of the other three datasets, in which the first three principal components of the kernel-based method are sometimes needed to separating the different categories or scenarios (Figures S5, S7 and S8, while only the first two principal components of the tree shape statistics considered in this paper are usually sufficient (Figures S6 and S9). However, both methods are clearly able to identify enough signal in the data to separate the categories or scenarios.

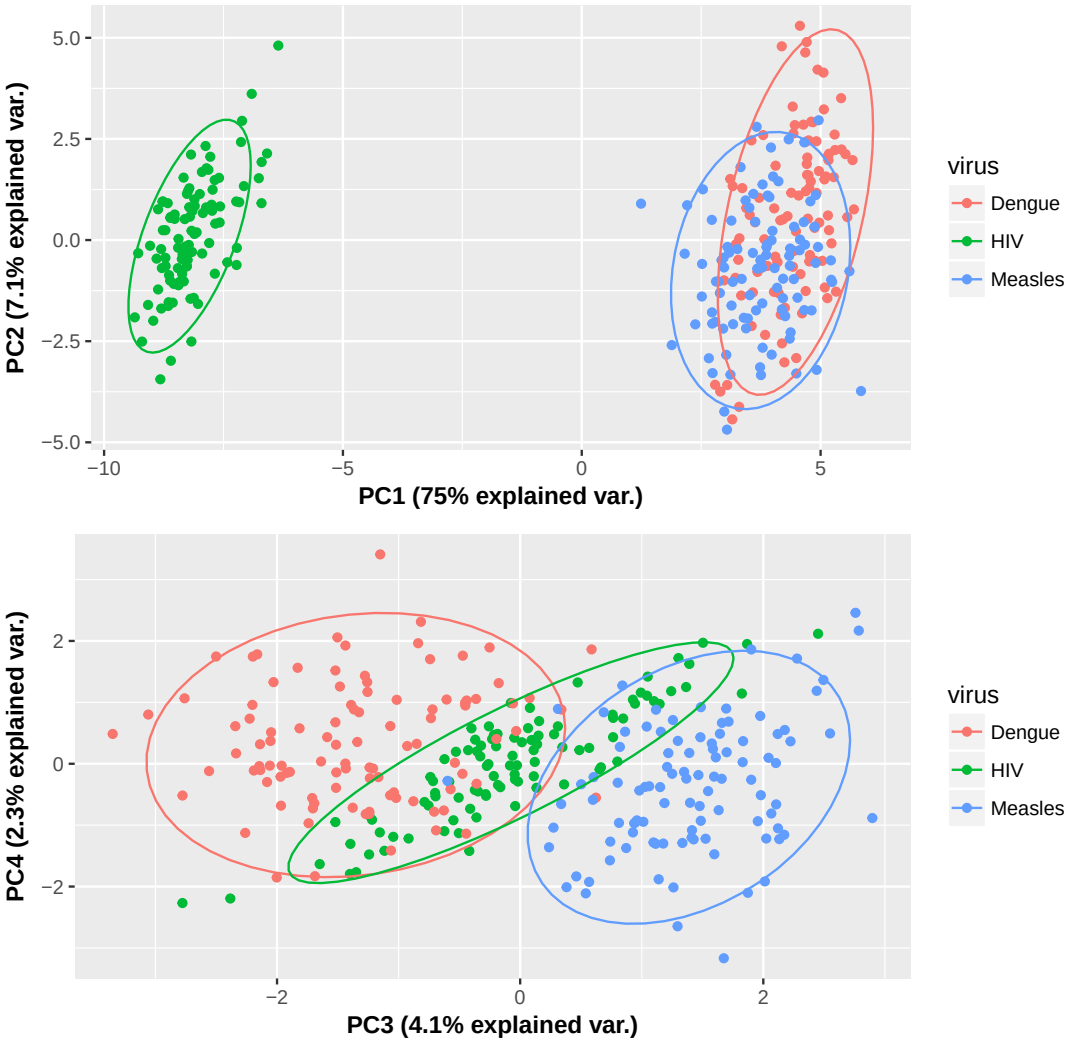


Figure S1: Kernel PCA plots of first four principal components illustrating the separation of HIV (green), Dengue (red) and Measles (blue) virus phylogenies by a tree kernel method.

6.4 Calculating the (degree) assortativity of a phylogenetic tree

Consider a phylogenetic tree on n tips as a graph. Its degree assortativity is the Pearson correlation between the degrees of the sources (heads) and targets (tails) of its edges. We direct each edge away from the root, towards the tips. We show below that for $n \geq 4$, only two values of the assortativity coefficient can occur (only one such value can obviously occur for $n \leq 3$).

We make use of the following formula for the Pearson correlation between $\vec{x} \in \mathbb{R}^N$ and $\vec{y} \in \mathbb{R}^N$:

$$r_{xy} := \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \sqrt{N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}.$$

Based on this formula, we note that only two options can occur; the first option is when the root's children are both internal nodes, and the second option is when one of its children is a tip and the other, an internal node (which occurs when, for instance, there is an outgroup in the data). In the first case, we use

$$\vec{x} = (\underbrace{3, \ldots, 3}_{n \text{ times}}; 2, 2; \underbrace{3, \ldots, 3}_{n-4 \text{ times}})$$
$$\vec{y} = (\underbrace{1, \ldots, 1}_{n \text{ times}}; 3, 3; \underbrace{3, \ldots, 3}_{n-4 \text{ times}})$$

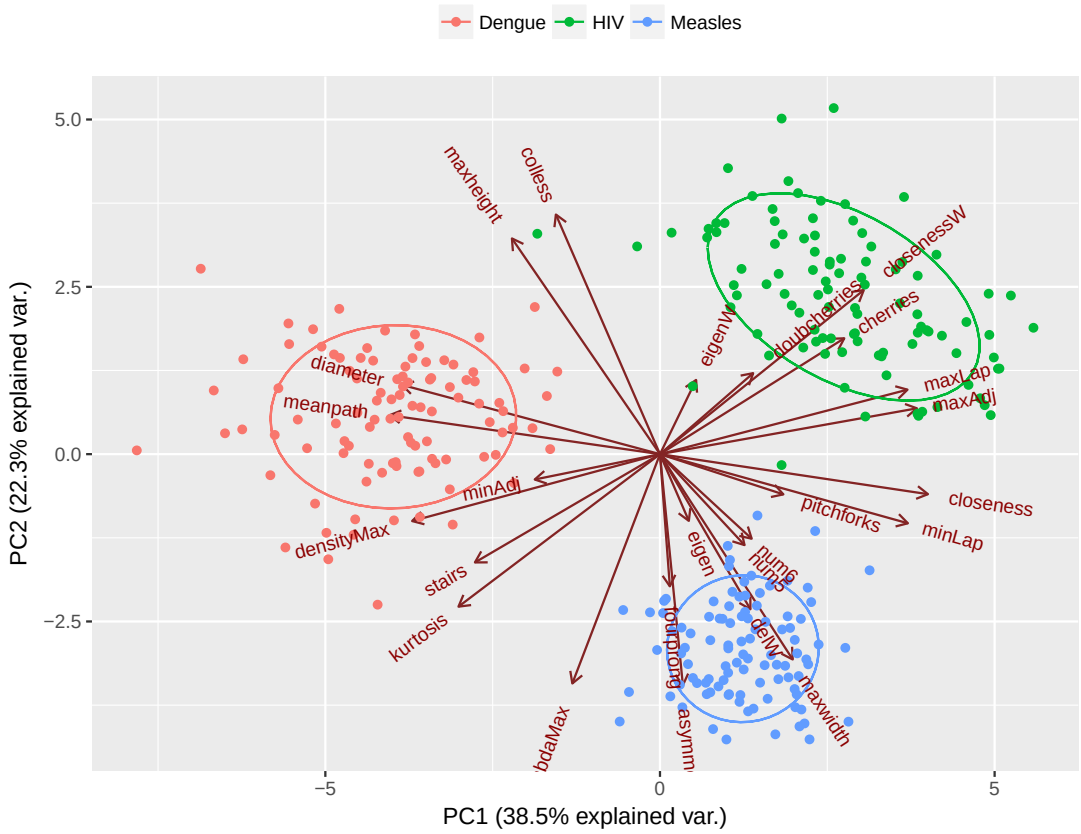


Figure S2: PCA biplots of first two principal components illustrating the separation of HIV (green), Dengue (red) and Measles (blue) virus phylogenies by the shape statistics.

since there are exactly n edges going from an internal node to a tip, 2 edges from the root to an internal node, and $n-4$ edges from an internal node to another internal node. An easy calculation then yields

$$A_1(n) := \frac{-\sqrt{n}}{n-2} \text{ and } A_1'(n) := \frac{-(n^2+2n+1)}{3n^2-4n-3},$$

where A_1 denotes the *directed* degree assortativity, calculated by substituting $\vec{x}$ and $\vec{y}$ directly into the formula above, while A_1' denotes the *undirected* degree assortativity, calculated by substituting $[\vec{x}\vec{y}]$ and $[\vec{y}\vec{x}]$ into it, where $[\vec{u}\vec{v}]$ denotes the concatenation of two vectors. The latter occurs because in the undirected case, each directed edge (ij) is counted twice, once as an edge from i to j and once as an edge from j to i.

In the second case, where the root has a tip as one of its children, the degree sequences are

$$\vec{x} = (\underbrace{3,\ldots,3}_{n-1 \text{ times}}; 2, 2; \underbrace{3,\ldots,3}_{n-3 \text{ times}})$$
$$\vec{y} = (\underbrace{1,\ldots,1}_{n-1 \text{ times}}; 1, 3; \underbrace{3,\ldots,3}_{n-3 \text{ times}})$$

and the corresponding values are

$$A_2(n) := \frac{-1}{\sqrt{n}(n-2)} \text{ and } A_2'(n) := \frac{-(n^2-2n+5)}{3n^2-4n-3}.$$

We note that $\lim_{n\to\infty} A_1(n) = 0 = \lim_{n\to\infty} A_2(n)$ and $\lim_{n\to\infty} A_1'(n) = -1/3 = \lim_{n\to\infty} A_2'(n)$.

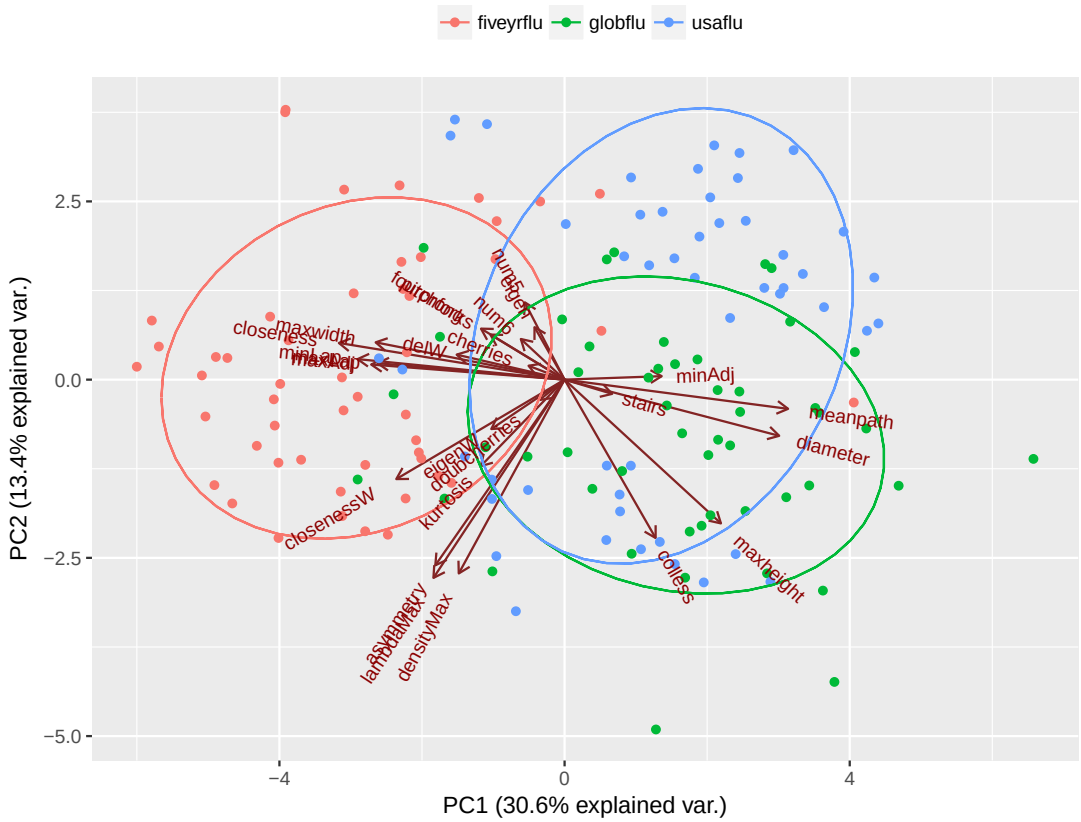


Figure S3: PCA biplots illustrating the separation among influenza A virus phylogenies sampled at global (green), five-year (red) and regional (United States, blue) scopes by using tree shape statistics

6.5 The distribution of the diameter and the Wiener index of trees

The distribution of diameters for $n = 23$ tips ranges from 8 to 23, with a mean of 14.147. Figure S10 shows the distribution of all the diameters, together with one of the trees achieving the minimum and maximum values of the diameter, respectively.

The distribution of Wiener indices for $n = 23$ tips ranges from 5382 to 8778, with a mean of 6516.541. Figure S11 shows the distribution of all the diameters, together with the trees achieving the minimum and maximum values of the Wiener index, respectively.

Figure S12 illustrates all 6 phylogenetic trees on $n = 6$ tips with their Wiener index and diameter, arranged in increasing order of the Wiener index, with ties broken by the diameter.

6.6 The distribution of betweenness/closeness/eigenvector centrality

The distribution of maximum betweenness centrality values for $n = 23$ tips ranges from 505 to 645, with a mean of 587.592. Figure S13 shows the distribution of all the maximum betweenness centrality values, together with one of the trees achieving the minimum and maximum values, respectively.

The distribution of maximum closeness centrality values for $n = 23$ tips ranges from 143 to 285, with a mean of 189.920. Figure S14 shows the distribution of all the maximum closeness centrality values, together with the trees achieving the minimum and maximum values, respectively.

The distribution of maximum eigenvector centrality values for $n = 23$ tips ranges from 0.2324 to 0.4336, with a mean of 0.3573. Figure S15 shows the distribution of all the maximum eigen-

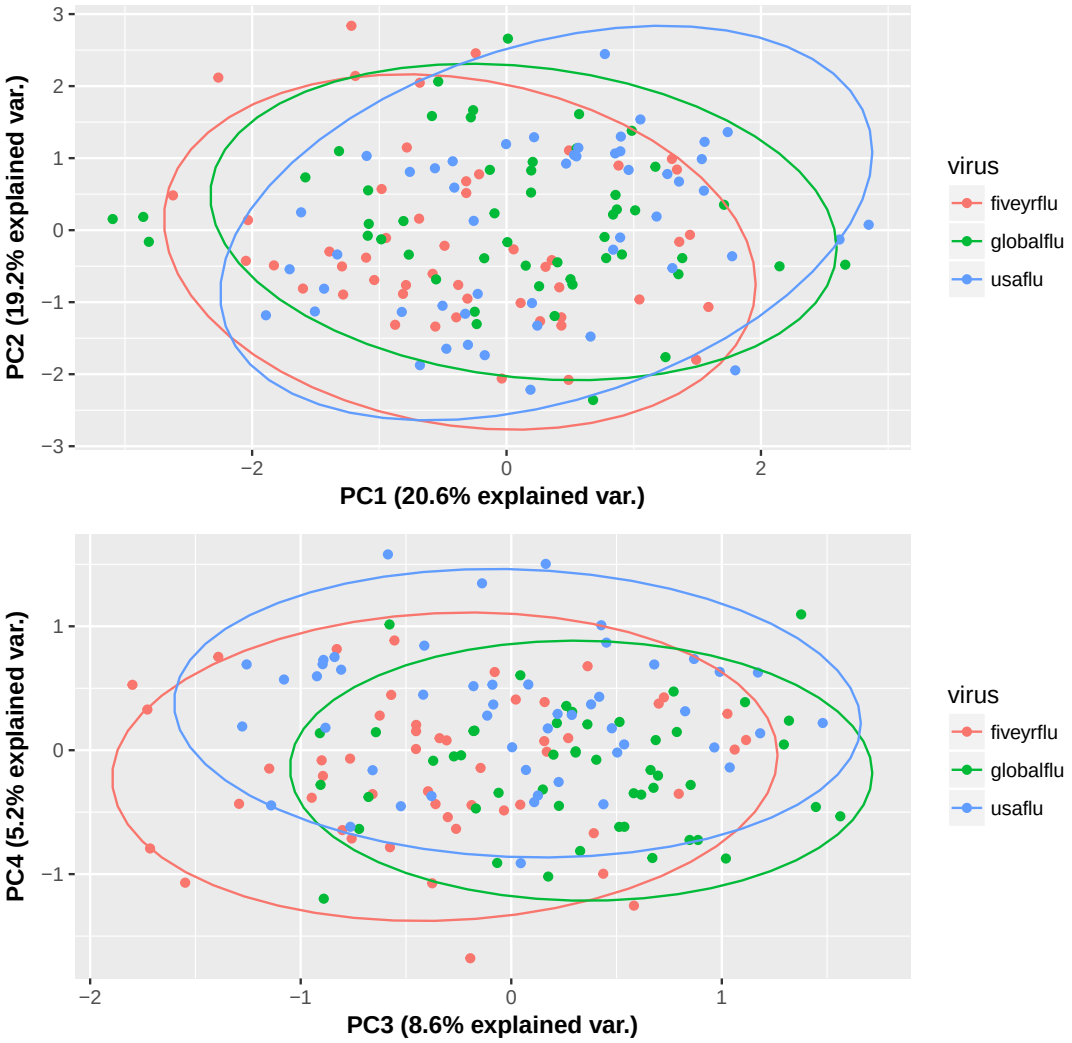


Figure S4: Kernel PCA plots illustrating the overall lack of separation among influenza A virus phylogenies sampled at global (green), five-year (red) and regional (United States, blue) scopes by using the tree kernel method.

vector centrality values, together with the trees achieving the minimum and maximum values, respectively.

Figure S16 illustrates all 6 phylogenetic trees on $n = 6$ tips with their betweenness, closeness, and eigenvector centrality, arranged in increasing order of the betweenness, with ties broken by the closeness.

6.7 Proof of the location of the maximum values of centralities

We now present a simple proof for the following facts, mentioned in the main text:

a. Betweenness centrality is maximized at the **tricenter** of the tree (the node which, if the tree is rerooted at it, has 3 descendant clades of sizes within one of each other), if one exists, and minimum at the tips

b. Closeness centrality is maximized at the **centroid** of the tree (the node which, if the tree is rerooted at it, has all descendant clades containing fewer than half the total nodes), and minimum at one of the tips

c. Eigenvector centrality is maximized at one of the internal nodes

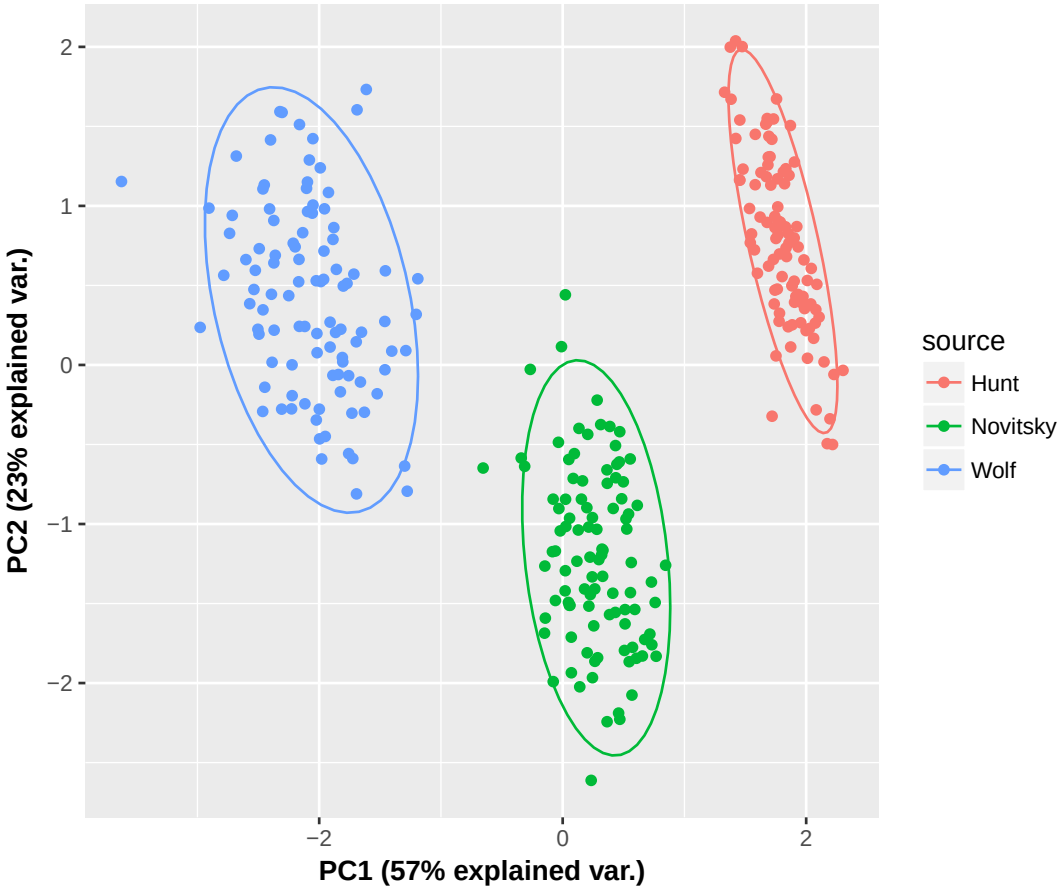


Figure S5: Kernel PCA plots illustrating the separation among HIV phylogenies sampled from a concentrated epidemic (blue), a generalized epidemic in a country (red) and in a village (green) by using the kernel method.

Proof:

a. Let v be any node in a tree T. If n_1 and n_2 are the sizes of the left and right subtrees of v and n_0 is the number of nodes outside T_v, the clade of T subtended by v, then its betweenness centrality is $n_0n_1 + n_0n_2 + n_1n_2$. We maximize this quantity subject to the constraint $n_0+n_1+n_2 = N-1$ and keep in mind that all of the sizes are integers. First, if $N \equiv 1 \mod 3$, all three of these quantities can be equal to $\frac{N-1}{3}$, and this matches the best we can do even if we allow fractional values. Second, if $N \equiv 2 \mod 3$, say $N = 3K+2$, the best option is to set one of the three values to $K+1 = \frac{N+1}{3}$ and make the remaining two equal to $K = \frac{N-2}{3}$. Finally, if $N \equiv 0 \mod 3$, say $N = 3K$, the best option is to set two of the three values to $K = \frac{N}{3}$ and make the remaining one equal to $K-1 = \frac{N}{3}-1$.

   In summary, the maximum is attained when $|n_0-n_1| \leq 1, |n_0-n_2| \leq 1, |n_1-n_2| \leq 1$. We call a node which satisfies this condition a **tricenter** of the tree. Not all trees have one - for instance, the complete binary tree on $n = 4$ tips does not have one. The uniqueness of a tricenter follows from the uniqueness of the centroid, proven as part of the next result. Furthermore, the value of betweenness centrality at a tricenter is close to $\frac{N^2}{3}$, while having $n_0 = 0$ (as would be the case for the root) results in a value of at most $\frac{N^2}{4}$, which is clearly suboptimal. Lastly, it is also clear that $n_1 = n_2 = 0$ (as would be the case for any tip) results in a betweenness centrality value of 0, the minimum possible.

b. Let u be the node of a phylogenetic tree T_0 that has the smallest farness; we will show that u must be a centroid, and conclude by showing that every phylogenetic tree has a unique

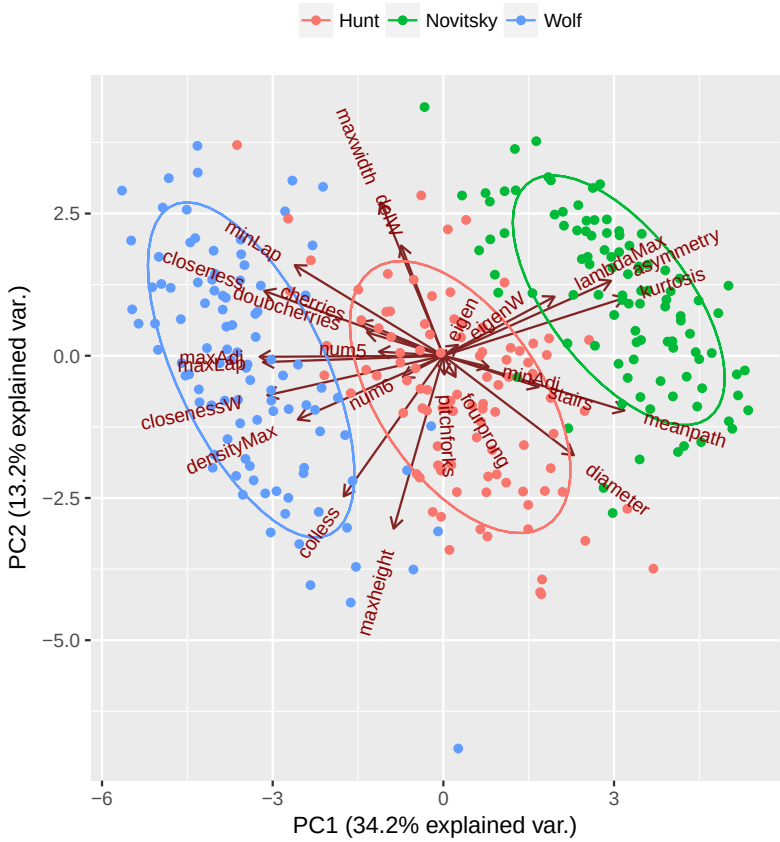


Figure S6: PCA biplots illustrating the separation among HIV phylogenies sampled from by using tree shape statistics.

centroid. For the first part, let us consider the tree T (which will not be binary if u is not the root) obtained by rerooting T_0 at u. Note that the distances in T are equal to the distances in T_0. To make things general we consider a tree with branch lengths (weights).

Let v be any child of u in T. Note that $d(v,x) = d(u,x) - w(uv)$ for any node x in the subtree T_v rooted at v, and $d(v,x) = d(u,x) + w(uv)$ for any node x outside T_v (including u itself). If we denote by $F(x)$ the farness of a node x, then

$$F(v) = \sum_{x \in T_v} (d(u,x) - w(uv)) + \sum_{x \notin T_v} (d(u,x) + w(uv))$$
$$= \sum_{x \in V(T)} d(u,x) + (N - 2|T_v|)w(uv) = F(u) + (N - 2|T_v|)w(uv).$$

This recurrence is in fact the basis for a linear-time algorithm for computing $F(x)$ for all nodes x in T_0, starting from the root. Recall that $w(uv) > 0$ by assumption. Therefore, in order for $F(u)$ to be the smallest farness, we must have

$$F(v) \geq F(u) \iff N - 2|T_v| \geq 0 \iff T_v \leq \frac{N-1}{2}, \tag{1}$$

since $N = 2n - 1$ is odd, and in fact, the inequalities must be strict. It follows that the subtree T_v of T rooted at any child v of u must contain fewer than half of all the nodes; therefore, u must satisfy the definition of a centroid. We note that two closely related results appear in the work on 1-medians of tree networks [21].

It remains to show that the centroid exists and is unique in a phylogenetic tree. For the existence, we apply a result by Camille Jordan [29] that proves the existence of a *separator*

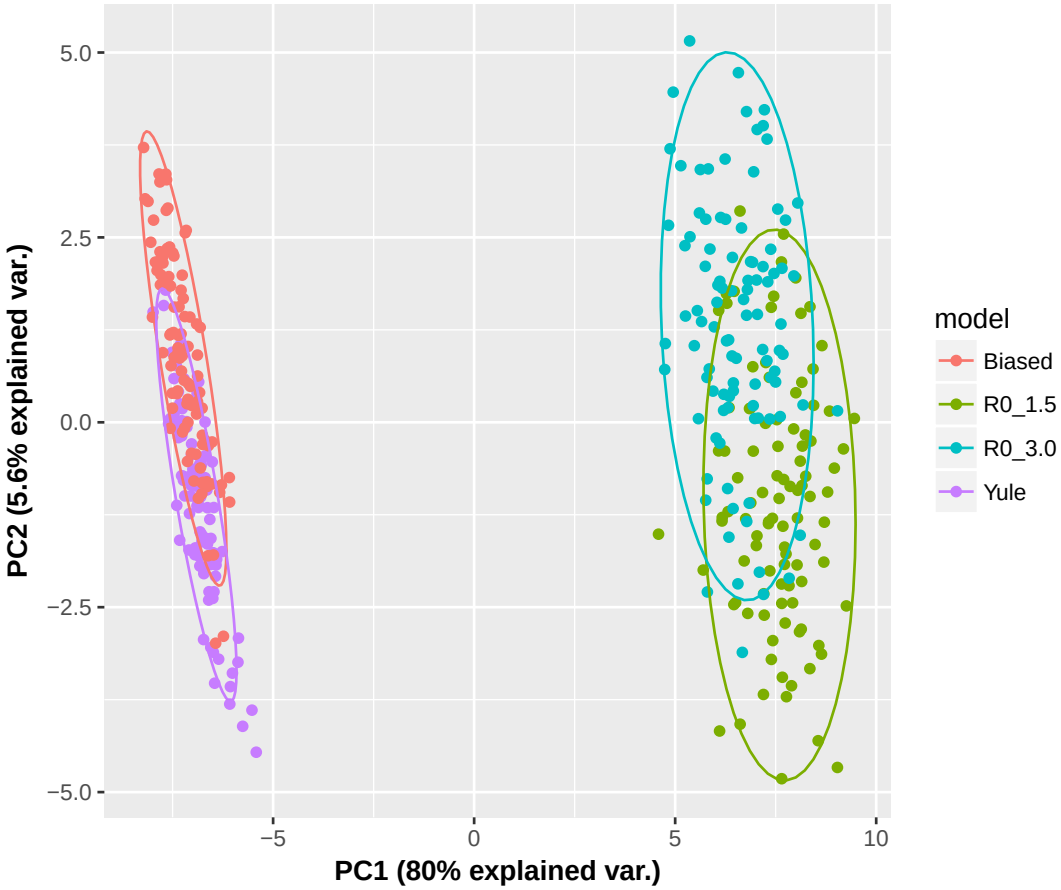


Figure S7: Kernel PCA plot illustrating the separation among phylogenies simulated from a biased model (red), a Yule model (purple) and two birth-death models (green and blue) by using the kernel method.

node v in any (not necessarily binary) tree T, which is a node defined by the property that any connected component obtained by removing v from T has size at most $\frac{N}{2}$, where N is the number of nodes in T. The reason that such a separator node is a centroid is that the connected components resulting from removing v are identical with the subtrees rooted at v if T is rerooted at v, and that N is odd, so $x \leq \frac{N}{2} \iff x < \frac{N}{2}$ for $x \in \mathbb{N}$.

For the uniqueness (in phylogenetic trees), suppose that this there is another node in T_0, say $w \neq u$, that also has the defining property of a centroid. Consider the connected component C^u of T_0 after w is removed that contains u, and the connected component C^w of T_0 after u is removed that contains w. Since u and w are centroids, C^u and C^w each contain fewer than half the nodes, so their union does not contain some node v of T_0. But this cannot happen because if $d(u,v) \leq d(w,v)$ then $v \in C^u$, while if $d(w,v) \leq d(u,v)$ then $v \in C^w$.

Lastly, we note that for any internal node v, there is some tip w such that the path from v to w strictly increases in farness; indeed, it is sufficient to always pick the child of the current node whose subtree contains fewer than half of all the nodes, by equation (1). Therefore, the maximum farness is attained at a tip, not at an internal node.

To connect the concepts of centroid and tricenter, we note that a tricenter, if one exists, is also a centroid, by definition; in particular, a tricenter, if it exists, is always unique; however, a phylogenetic tree always has a centroid, while it may not have a tricenter.

c. Recall the assumption that the tree T has arbitrary positive branch lengths. Let v be a tip

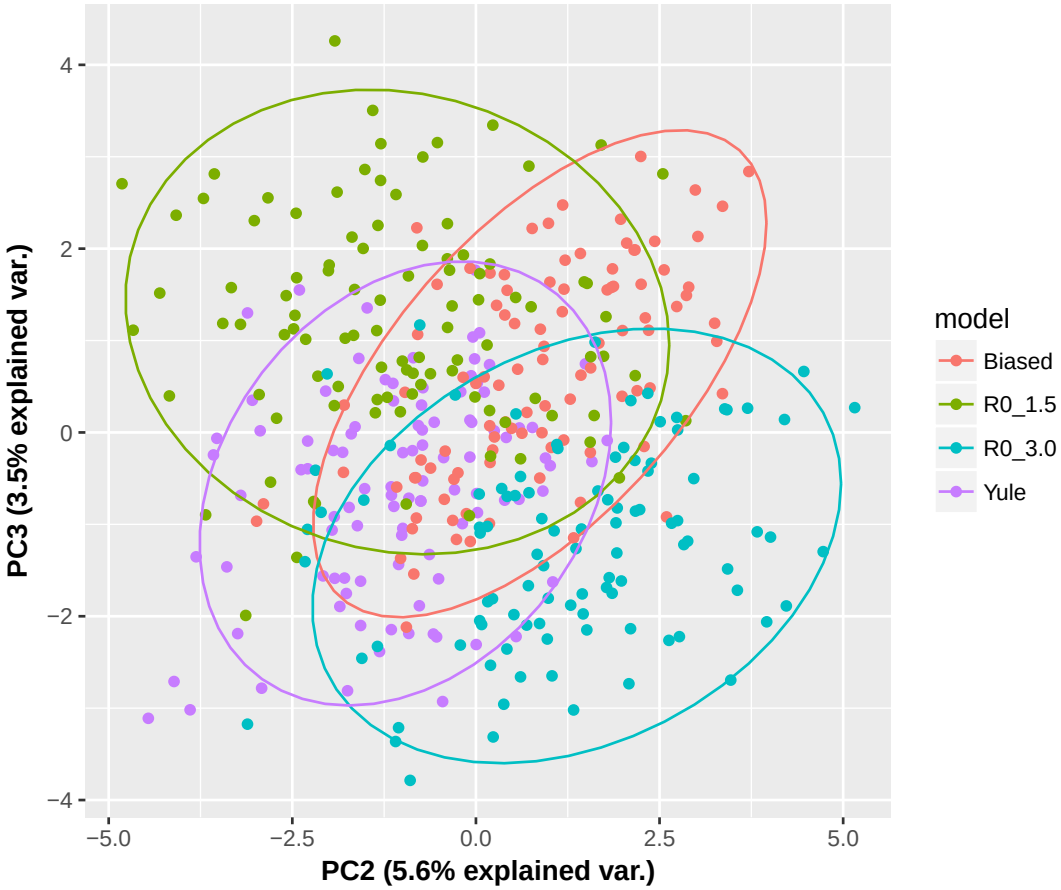


Figure S8: The same kernel PCA plot as above, components 2 and 3.

with a parent u. The defining equations for the Perron-Frobenius eigenvector specialized to the entries corresponding to u and v say that

$$\lambda e_v = e_u w(uv);$$
$$\lambda e_u = \sum_{w \in \Gamma(u)} e_w w(uw) > e_v w(uv)$$

where the inequality follows from the fact that u has at least one other neighbor, and all the branch lengths and entries of $\vec{e}$ are positive. By cross-multiplying the inequality with the equality, we get

$$\lambda e_u(e_u w(uv)) > e_v w(uv)(\lambda e_v) \implies e_u^2 > e_v^2 \implies e_u > e_v,$$

since $\lambda > 0$ and $w(uv) > 0$. Hence, a tip cannot contain the maximum eigenvector centrality, which is therefore attained at an internal node.

6.8 The distance Laplacian spectrum does not uniquely define a tree

We give an example of a pair of trees on $n = 3$ tips, with different branch lengths, whose distance Laplacian matrices are co-spectral. We obtained them by exploring all possible integer branch lengths, in order of total branch length (from smallest to largest), and used Maple [46] to check, for each resulting weighted topology, the existence of a different set of branch lengths that would result in a distance Laplacian matrix having the exact same characteristic polynomial, and hence the exact same spectrum. The first tree is therefore the tree with integer branch lengths having

Figure S9: PCA biplots illustrating the separation among phylogenies simulated from a biased model (red), a Yule model (purple) and two birth-death models (green and blue) by using tree shape statistics.

the smallest possible total length and admitting a co-spectral tree, the second tree, with respect to the distance Laplacian matrix.

The first tree has branch lengths $a = 2$ from the root to the first tip, $b = 2$ from the root to the internal node, and $c = d = 1$ from the internal node to the remaining two tips. The second tree has corresponding branch lengths $a \approx 0.16754417767858494454487$, $b \approx 3.5136399160176469472222$, $c \approx 1.3164466748555757609955$, $d \approx 0.24554927343936887369 27$ (the exact expressions involve the roots of a polynomial of degree 22 and are provided in the Maple worksheet as a **Supplementary File**). These results were verified using the *RPANDA* package [48] in R.

6.9 Distance Laplacian spectra do determine small unweighted trees

We created a computer program to exhaustively test whether any pair of distinct phylogenetic trees had identical distance Laplacian spectra. The trees were generated by the method proposed by Gang Li in his Masters thesis [37], which only requires enough space for a single tree and a constant amount of operations per tree generated. For each tree we computed the distance Laplacian spectrum and compared it to that of all the other trees of the same size. Since the number of unlabelled phylogenetic trees is known to grow as $C\alpha^n n^{-3/2}$, where $C \approx 0.792$ and $\alpha \approx 2.483$ [37], it quickly becomes infeasible to enumerate and keep in memory the full spectra of the distance Laplacian matrix for each possible tree. For this reason we made use of several optimizations.

First, we observed that two trees can only have cospectral distance Laplacian matrices if the traces of these matrices, known as the Wiener indices of the trees (the sum of all node-node

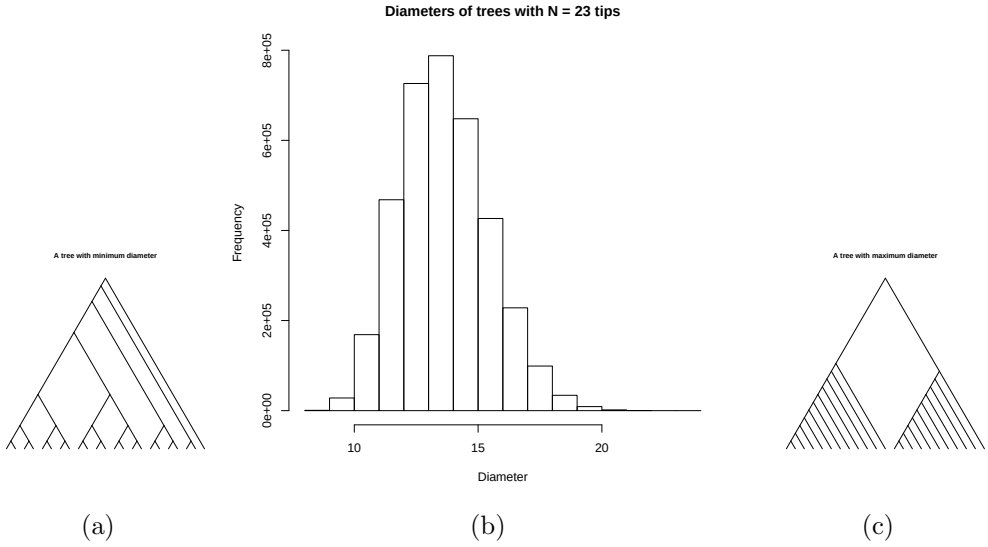
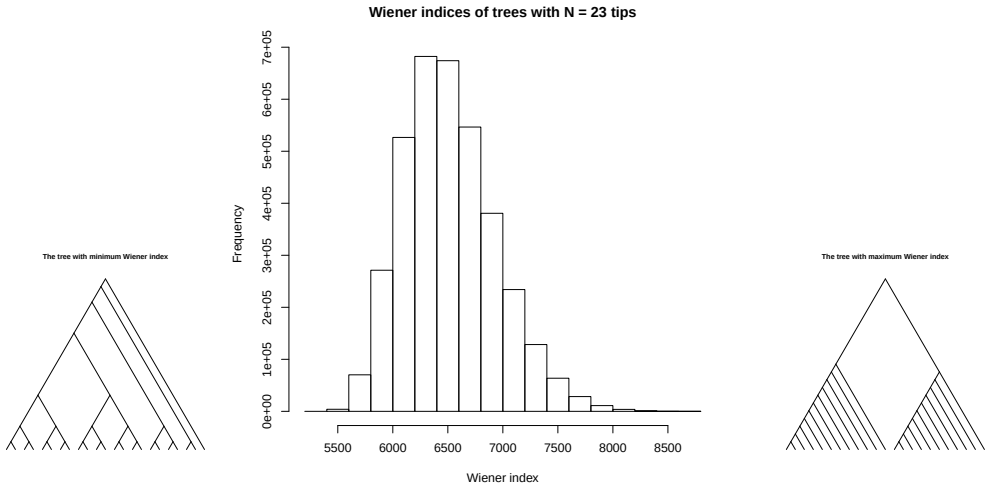


Figure S10: The distribution of diameters for phylogenetic trees on $n = 23$ tips, with two extremal trees

distances) [45], are equal. In particular, this means that the computation could be parallelized if the values of the possible Wiener indices were known in advance, together with their multiplicities. Second, we observed that the distribution of the Wiener indices of a tree on n tips could be pre-computed efficiently from the heights and Wiener indices of the subtrees composing it. Lastly, we were able to augment the tree generation algorithm with an additional set of instructions that allowed us to also compute the Wiener index while retaining the constant amortized time complexity.

In this manner, we pre-computed the distribution of the Wiener indices, decided on a split of the trees between processes, and limited the spectral comparisons to those within the same process. We performed the comparisons themselves in two passes; in the first pass, all trees with appropriate Wiener index were generated and their highest and lowest eigenvalue stored in a large matrix; in the second pass, only those trees that had at least one match were re-generated, and their full spectra compared. The computation for $n = 32$ tips, involving 7.86 billion trees split between 10 processors, required approximately 2 weeks of computation on a cluster, and did not produce any co-spectral trees, although some of the spectra were within 0.01 distance in the max norm (L_∞).



(a) (b) (c)

Figure S11: The distribution of Wiener indices for phylogenetic trees on $n = 23$ tips, with two extremal trees

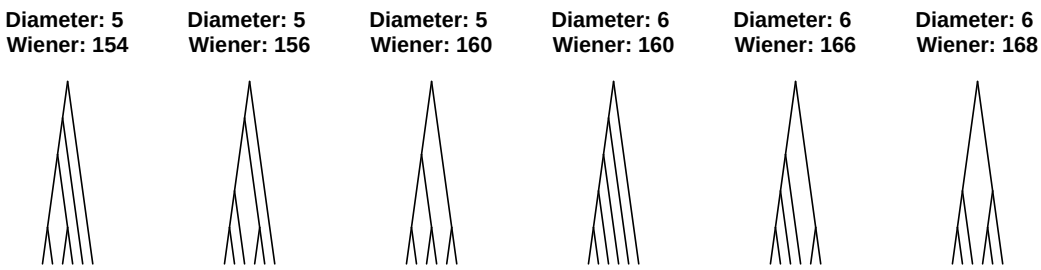
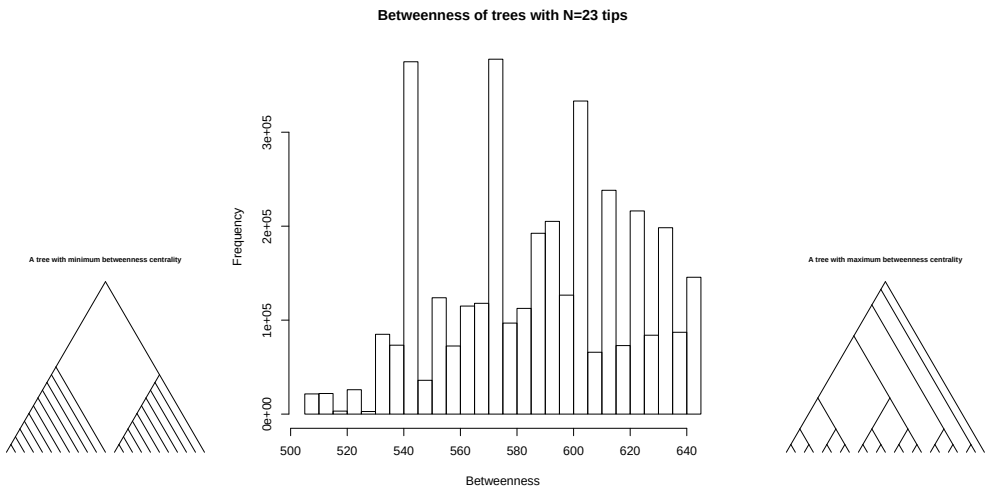
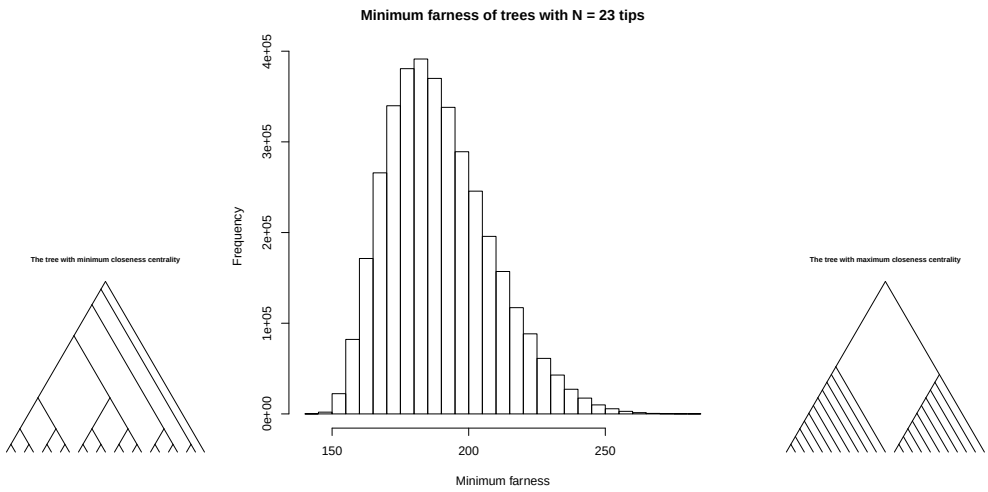


Figure S12: The 6 phylogenetic trees on $n = 6$ tips, with their diameter and Wiener index



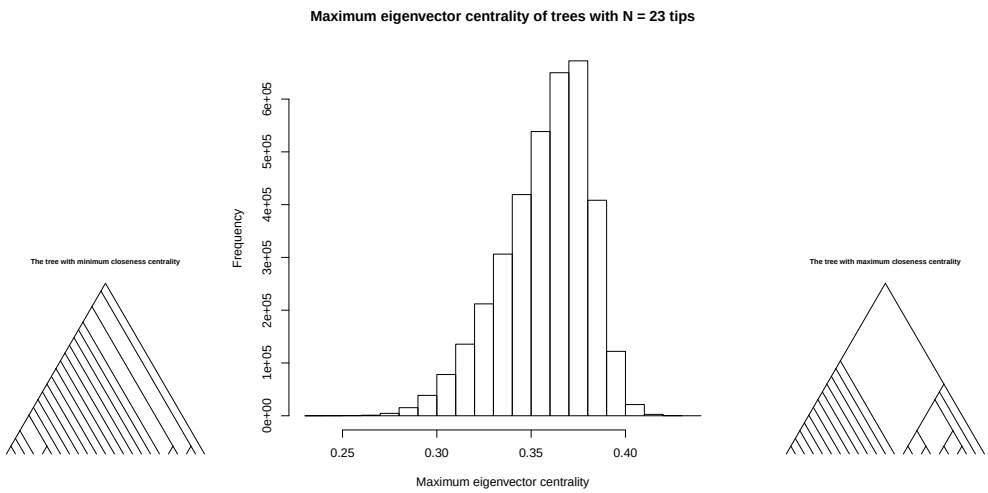
(a) (b) (c)

Figure S13: The distribution of maximum betweenness centralities for phylogenetic trees on $n = 23$ tips, with two extremal trees



(a) (b) (c)

Figure S14: The distribution of minimum farness (inverse of maximum closeness centralities) for phylogenetic trees on $n = 23$ tips, with two extremal trees



(a) (b) (c)

Figure S15: The distribution of maximum eigenvector centralities for phylogenetic trees on $n = 23$ tips, with two extremal trees

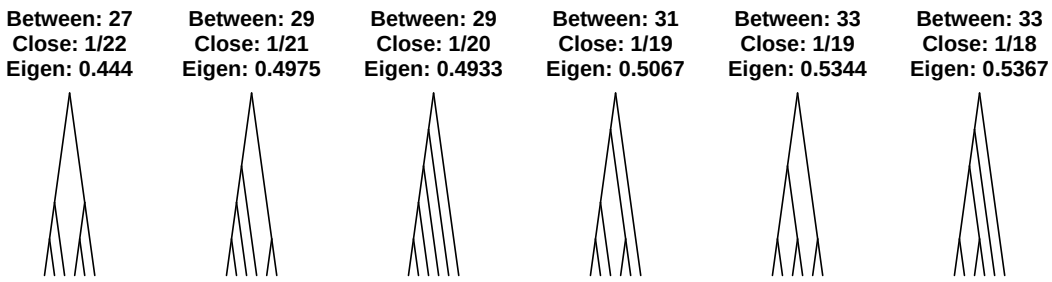


Figure S16: The 6 phylogenetic trees on $n = 6$ tips, with their betweenness, closeness, and eigenvector centrality