

# A statistical framework for comparing epidemic forests

Cyril Geismar<sup>1,2,3</sup>, Peter J. White<sup>2,3,4</sup>, Anne Cori<sup>2,3, \*</sup>, and Thibaut Jombart<sup>2,3, \*</sup>

<sup>1</sup>Bloomberg School of Public Health, Johns Hopkins University,  
Baltimore, United States

<sup>2</sup>MRC Centre for Global Infectious Disease Analysis, Imperial  
College School of Public Health, London, United Kingdom

<sup>3</sup>NIHR Health Protection Research Unit in Modelling and Health  
Economics, Imperial College School of Public Health, London,  
United Kingdom

<sup>4</sup>UK Health Security Agency, London, United Kingdom  
\* authors contributed equally

## 1 Abstract

Inferring who infected whom in an outbreak is essential for characterising transmission dynamics and guiding public health interventions. However, this task is challenging due to limited surveillance data and the complexity of immunological and social interactions. Instead of a single definitive transmission tree, epidemiologists often consider multiple plausible trees forming *epidemic forests*. Various inference methods and assumptions can yield different epidemic forests, yet no formal test exists to assess whether these differences are statistically significant. We propose such a framework using a chi-square test and permutational multivariate analysis of variance (PERMANOVA). We assessed each method's ability to distinguish simulated epidemic forests generated under different offspring distributions. While both methods achieved perfect specificity for forests with 100+ trees, PERMANOVA consistently outperformed the chi-square test in sensitivity across all epidemic and forest sizes. Implemented in the R package *mixture*, we provide the first statistical framework to robustly compare epidemic forests.

## 2 Author Summary

Identifying who infected whom is a central part of outbreak investigation. It helps trace the source of infection, uncover missing cases, identify superspreaders, and describe broader dynamics of transmission such as its speed, pattern, and scale. With the advent of pathogen sequencing and digital contact tracing, computational models have become the standard approach for reconstructing outbreaks. These probabilistic models do not identify a single definitive history of who infected whom (*i.e.* a transmission tree), but a collection of plausible alternatives, which we call ‘epidemic forests’. Different modeling assumptions or data sources can produce different epidemic forests, but until now, there has been no formal way to determine whether these differences are meaningful.

We present the first statistical framework designed to compare epidemic forests. We evaluate two methods: one that counts how often specific transmission pairs appear, and another that compares the structure of transmission trees. Testing these methods on simulated outbreaks, we found that both successfully identified when forests represented identical transmission dynamics, but one method outperformed the other in identifying forests representing distinct transmission dynamics. Our framework, implemented in the R package `mixtree`, enables epidemiologists to validate and compare outbreak reconstruction approaches, supporting more reliable investigations.

### 3 Introduction

Tracking who infected whom is central to outbreak investigations. Transmission trees, modelled as directed acyclic graphs (DAGs) where vertices represent infected individuals and directed edges indicate transmission events, delineate infector-infectee relationships [1]. These representations can assist epidemiologists in identifying introduction and superspreading events [2, 3], whilst also elucidating broader transmission dynamics relevant to outbreak response. The topology of transmission trees, defined by the arrangement of vertices and edges, encodes key epidemiological parameters. The out-degree distribution of vertices represents the number of secondary infections per infected individual (*i.e.* the offspring distribution), revealing the degree of heterogeneity in transmission [4–7]. Branching patterns inform on transmission dynamics between groups [8], revealing group reproduction numbers [5] and transmission patterns, for example, between healthcare workers and patients in nosocomial outbreaks [9] or between children and adults in schools [10].

The inference of transmission trees is challenging and often characterised by large uncertainty, partly due to the lack of discriminatory power in choosing between possible transmission pairs [5, 9, 11], incomplete surveillance data and diverse, sometimes conflicting sources (*e.g.* contact, temporal, spatial, or genetic data) [12]. Additional complexities arise from varying methodological approaches [12] and pathogen evolution mechanisms that are difficult to model (*e.g.* within-host evolution and transmission bottleneck [7, 13]). Consequently, outbreak reconstruction often yields *epidemic forests*, which are collections of plausible transmission trees rather than a singular definitive representation of who infected whom.

Without formal statistical methods to differentiate epidemic forests, determining whether differences between them represent meaningful variations in transmission dynamics or uncertainty in tree reconstruction is challenging. Such distinction would help validate convergence when repeated model runs produce statistically similar forests and assess whether competing inference approaches or alternative data sources yield significantly different forests.

Bayesian inference methods have emerged as the gold standard for transmission tree reconstruction, with various approaches differing in their assumptions, data requirements, and inference strategies [12]. In this context, epidemic forests represent samples drawn from a model’s posterior distribution of transmission trees. To assess the performance of the inference process, researchers rely on general Markov Chain Monte Carlo (MCMC) diagnostics, applied to scalar parameter chains rather than the inferred trees themselves. These diagnostics evaluate convergence through trace plot inspection and the Gelman-Rubin statistic [14], assess sampling efficiency through effective sample size calculations, and check model fit using posterior predictive checks [15]. In parallel, *consensus trees* are used to summarise epidemic forests, typically representing , for each

case, the infector with the highest posterior support across samples [11, 16–19]. However, these trees are often abstract representations rather than plausible transmission scenarios, potentially introducing cycles or multiple index cases. While algorithms such as Edmonds can enforce a valid tree topology, the resulting consensus tree may correspond to a combination of ancestries that was never observed as a complete tree in the posterior [19–21].

Consequently, standard MCMC diagnostics assess parameter chains rather than the inferred transmission events, while consensus trees ignore the uncertainty in who infected whom and may misrepresent key epidemiological features. This underscores the need for specialised statistical methods that can differentiate epidemic forests while accounting for uncertainty and relevant topological properties.

Here, we introduce a statistical framework for testing differences between epidemic forests. We consider two alternative methods: a chi-square ( $\chi^2$ ) test [22] used to compare the frequency of infector-infectee pairs between forests, and a permutation-based multivariate analysis of variance (PERMANOVA) [23] which compares topological distances between trees within and between forests. Both methods are summarised in Figure 1. We evaluated the performance of each method by comparing simulated epidemic forests with varying offspring distributions, measuring their ability to correctly identify forests stemming from distinct (sensitivity) or identical (specificity) generative processes (see Methods).

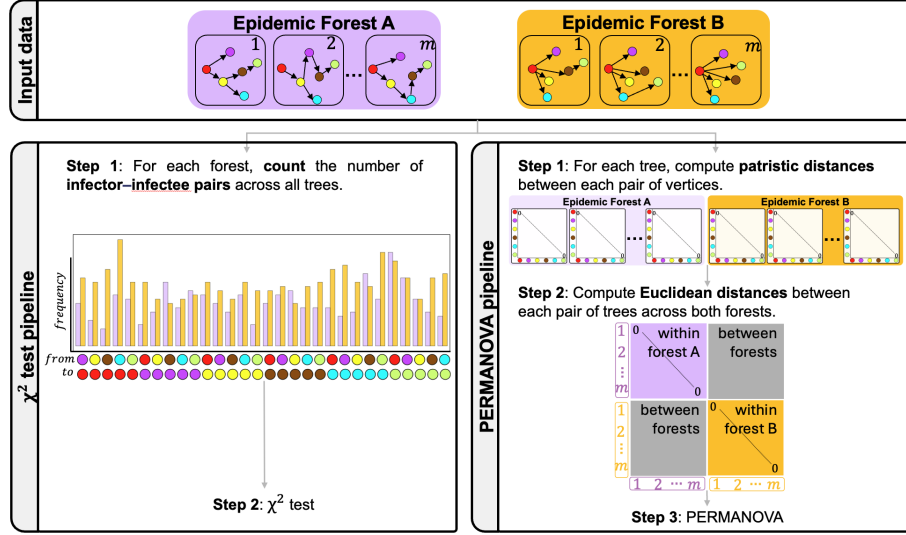


Figure 1: Statistical framework for comparing epidemic forests.

Diagram illustrating the methods for comparing epidemic forests  $A$  (pink, *e.g.* no superspreading) and  $B$  (orange, *e.g.* superspreading). Coloured dots represent infected cases, and arrows indicate transmission events. Left: The  $\chi^2$  test compares the frequency of infector-infectee pairs between forests. Right: The PERMANOVA method first calculates pairwise graph distances (number of transmission events between two cases, plus one - see Fig. S1) between vertices within each tree, converts these to a Euclidean distance matrix between all trees, and tests for significant topological differences between the forests using permutation-based testing. Both methods test the null hypothesis that the compared epidemic forests stem from the same generative process.

## 4 Results

We simulated pairs of epidemic forests that were either stemming from the same, or different generative processes (see Methods). We systematically varied epidemic size ( $\epsilon$ : 20–200 cases), forest size ( $m$ : 20–200 trees per forest), and the parameters of the negative binomial offspring distribution ( $R_0$ : 1.5–3 and  $k$ : 0.1–Poisson-like), which determine the mean and dispersion of secondary infections. We simulated 90,000 epidemic forests (see Methods) based on which we conducted 5,760,000 tests to measure sensitivity and specificity for the  $\chi^2$  test and PERMANOVA.

Overall results are presented as Receiver-Operator Characteristic (ROC) curves (supplementary Fig. S6), with area under the curve (AUC) provided in supplementary Figure S7. Both methods exhibited near perfect specificity ( $> 97\%$ ), *i.e.* the ability to correctly identify forests drawn from identical generative processes, across all epidemic or forest sizes (Fig. 2, row 4). The  $\chi^2$  test had a negligible advantage in specificity (+1.5%) when the number of trees in each forest was small ( $m \leq 50$ ). Across the aggregated simulation results, sensitivity was near perfect once the forest size reached 50–100 trees, with AUC nearing 1 (supplementary Fig. S6). However, these results varied substantially between methods and simulation settings.

The methods differed substantially in their sensitivity, *i.e.* their ability to correctly identify forests drawn from different generative processes, with PERMANOVA consistently outperforming the  $\chi^2$  test across all scenarios (Fig. 2, row 1–3). However, the magnitude of PERMANOVA’s advantage varied considerably depending on which parameters differed between forests. A logistic regression model explained 58% of the variance in test sensitivity (Pseudo  $R^2$  [24] = 0.58), with method choice, forest size, epidemic size, and differences in  $R_0$  and  $k$  as key predictors (Table 1). Compared to the  $\chi^2$  test, PERMANOVA showed much greater sensitivity when forests differed in their dispersion parameter ( $\Delta_k$ ), with 51-fold higher odds of correctly distinguishing overdispersed from Poisson-like forests ( $\Delta_{k_{(0,1] \text{ vs. Poisson}}}$ ), and 8-fold higher odds when comparing forests with different degrees of overdispersion ( $\Delta_{k_{(0,1] \text{ vs. } (0,1]}}$ ) (Table 1). When forests differed in dispersion and contained at least 100 trees, PERMANOVA achieved near-perfect sensitivity (98.7%), irrespective of epidemic size (Fig. 2, rows 1 and 3, column 3).

In contrast, PERMANOVA’s advantage over the  $\chi^2$  test diminished when forests differed only in reproduction number ( $OR = 3$ , Table 1). When both forests shared strong overdispersion (common  $k \leq 0.5$ ), high stochastic variability in individual transmission limited the ability to detect differences in  $R_0$  up to 1 ( $\Delta R_0 \leq 1$ ), yielding low sensitivity even with 200 trees per forest (52% across epidemic sizes; supplementary Figures S3 and S4). Sensitivity improved progressively as the common dispersion parameter approached Poisson-like transmission ( $k \rightarrow \infty$ ) or as epidemic size increased (supplementary Fig. S4).

In addition to higher sensitivity, PERMANOVA produced consistently narrower

p-value distributions than the  $\chi^2$  test, with interquartile ranges substantially smaller across all scenarios (supplementary Fig. S2).

Both methods' sensitivity increased with forest size (OR = 3, 6, 12 for  $m= 50$ , 100 and 200 respectively) but showed opposite correlations with epidemic size: PERMANOVA's sensitivity rose with larger epidemics (OR= 2, 4 and 5 for  $\epsilon= 50$ , 100, and 200 respectively), whereas the  $\chi^2$  test's sensitivity declined (OR= 0.5, 0.3, and 0.1) (Table 1, Fig. 2).

Our findings establish PERMANOVA as the superior method for comparing epidemic forests when sufficient samples are available ( $m \geq 100$ ), providing excellent sensitivity and specificity regardless of epidemic size.

Table 1: Logistic regression results for the sensitivity model (Eq.10)

Predictor		Odds Ratio	p-value
Intercept		0.01	<0.001
Method	PERMANOVA	1.94	<0.001
	50	2.93	<0.001
	100	6.10	<0.001
Forest size ( $m$ )	200	12.06	<0.001
	50	0.54	<0.001
	100	0.30	<0.001
Epidemic size ( $\varepsilon$ )	200	0.15	<0.001
Parameter difference ( $\Delta$ )	$R_0$	1.70	<0.001
	$k$		
	(0, 1] vs. (0, 1]	33.04	<0.001
	(0, 1] vs. Poisson	36.74	<0.001
	$R_0 : k$		
	(0, 1] vs. (0, 1]	0.62	<0.001
	(0, 1] vs. Poisson	0.61	<0.001
PERMANOVA : $\varepsilon$	50	4.19	<0.001
	100	11.85	<0.001
	200	33.94	<0.001
PERMANOVA : $\Delta$	$R_0$	3.01	<0.001
	$k$		
	(0, 1] vs. (0, 1]	7.77	<0.001
	(0, 1] vs. Poisson	51.38	<0.001
	$R_0 : k$		
	(0, 1] vs. (0, 1]	0.21	<0.001
	(0, 1] vs. Poisson	0.15	<0.001

*Note:* Pseudo  $R^2 = 0.58$  [24]. All p-values are < 0.001 due to the large simulation sample size ( $n = 5,040,000$ ). ‘.’ denotes the interaction term. The reference categories are:  $\chi^2$ test for method, 20 for  $\varepsilon$  and  $m$ , 0 for  $\Delta k$ .



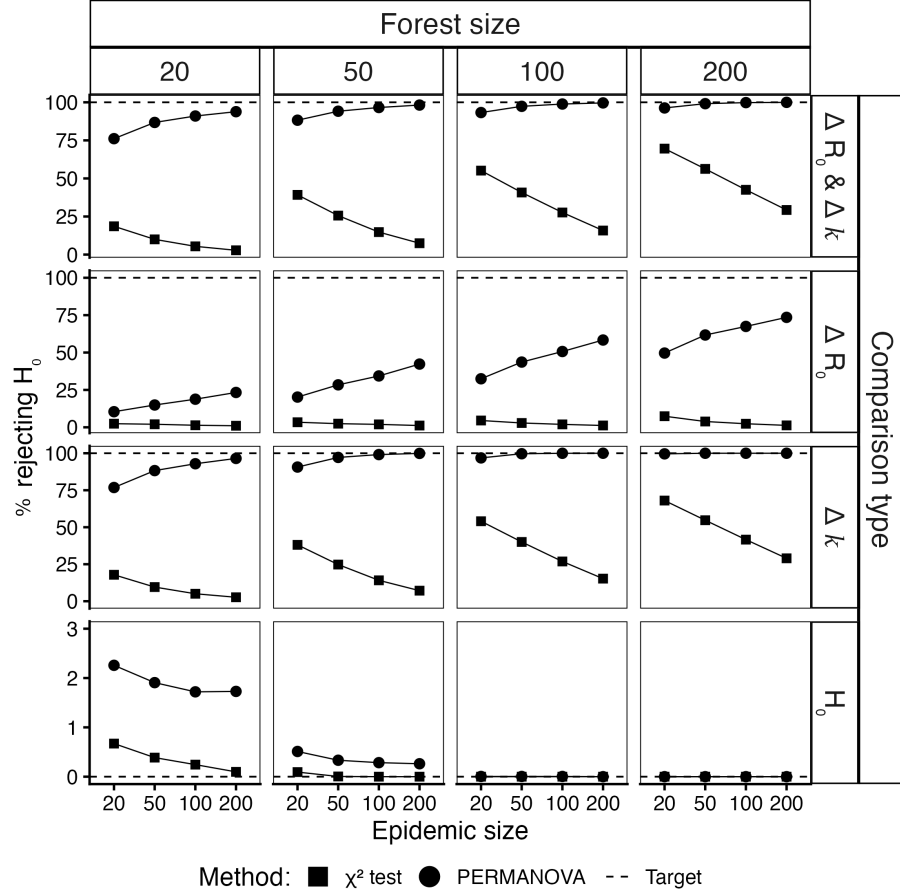


Figure 2: Performance of  $\chi^2$  and PERMANOVA for comparing epidemic forests.

This figure summarises simulation results for the  $\chi^2$  test (square points) and PERMANOVA (circular points) across varied parameter conditions. The y-axis shows the percentage of tests rejecting the null hypothesis ( $H_0$ ) of no difference between forests. The x-axis displays the epidemic sizes. Column panels refer to the forest size (*i.e.* the number of trees in each forest). Row panels refer to the type of differences between the two forests (with  $H_0$  for no differences).

## 5 Discussion

We evaluated two statistical approaches, the chi-square ( $\chi^2$ ) test and PERMANOVA, for distinguishing between collections of transmission trees (*i.e.* epidemic forests) originating from different generative processes, defined by the mean and dispersion of their offspring distribution (*i.e.* the distribution of secondary cases generated by each infected individual). The  $\chi^2$  test tests for differences in the frequency of infector-infectee pairs between epidemic forests, treating each pair as isolated edges without considering their relative position within each tree. In contrast, PERMANOVA leverages customised tree-based distance metrics to quantify meaningful epidemiological differences between tree topologies, which may better signal distinct pathogen transmission dynamics.

Our simulations showed that PERMANOVA consistently outperformed the  $\chi^2$  test in distinguishing epidemic forests generated under different offspring distributions. It achieved near-perfect sensitivity when forests differed in their dispersion parameter across all epidemic sizes (20-200 cases), provided forests contained at least 100 trees. However, its performance declined when forests differed solely in their mean reproduction number, especially for forests with high overdispersion (common  $k < 0.5$ ) (supplementary Fig. S4). In such settings, the substantial stochastic variability in individual transmission masked differences in mean transmissibility. Although the  $\chi^2$  test also demonstrated excellent specificity, its sensitivity was consistently lower across all scenarios and declined further as epidemic size increased. Larger epidemics produced higher forest entropy, which indicates greater variation in who infected whom across trees (supplementary Fig.S5, TableS1). Increased entropy yielded sparse contingency tables with many low expected counts and growing degrees of freedom, which reduced the statistical power of the  $\chi^2$  test (see Methods, Eq.2). In contrast, PERMANOVA became more sensitive as epidemics grew given that additional transmission events reduced the variance in within group distances, increasing the F-statistic (see Methods, Eq.9).

Computationally, both methods scale with epidemic size, although PERMANOVA incurs greater computational expense (see supplementary material). Parallelisation and constrained permutation (for PERMANOVA [25]) or replicates used in the Monte Carlo test (for  $\chi^2$  test [26]) make both methods applicable to most contexts. When comparing two forests, each with 100 trees and 100 vertices, the  $\chi^2$  test takes 0.5 seconds, while PERMANOVA takes an average of 5 seconds (supplementary material Table S2).

To facilitate accessibility of these methods, we have developed *mixtree* [27], a free, open-source R package available on CRAN [26]. *mixtree* implements both the  $\chi^2$  test and PERMANOVA methods described in this study.

The proposed framework addresses several needs for outbreak reconstruction. First, it provides a formal approach for assessing MCMC convergence in ‘tree space’ by comparing epidemic forests sampled from independent MCMC chains, which should be statistically indistinguishable when converged. This method complements existing diagnostics that focus on scalar parameter chains,

which do not fully capture the complex tree structures that form the primary output of Bayesian inference models. Second, it enables rigorous comparison between competing models with different assumptions about transmission dynamics, facilitating evidence-based model selection. Third, it can detect whether incorporating additional data sources (*e.g.* contact tracing [28]) into reconstruction efforts significantly alters the resulting transmission trees, helping researchers evaluate the value of supplementary data. However, it cannot independently determine which reconstruction is more accurate without additional validation measures.

Our study focused on comparing two forests of equal size for computational feasibility. However, both methods can compare any number of forests of varying sizes sharing the same set of vertices, as implemented in our *mixtree* package. Nonetheless, the two methods do not share identical limitations. PERMANOVA assumes full graph connectivity [29], so it cannot accommodate multiple introductions that result in disconnected trees. In contrast, the  $\chi^2$  test can handle multiple introductions by assigning them to a dedicated category in the edge list. In the presence of unobserved cases, the  $\chi^2$  test cannot distinguish between direct and unobserved intermediate transmissions. Importantly, PERMANOVA could be extended by modelling epidemiological, spatial or genetic distances as edge weights. For example, these weights could represent the number of infection generations between pairs of cases, thus accounting for unobserved cases. The simple graph distance used here could be replaced with a more complex metric that incorporates additional edge characteristics (*i.e.* weights) such as the number of generations between observed cases, or the time difference between their symptoms [30] or infection dates. While our simulation framework assessed method performance when the forest’s generative process differed only in its offspring distribution, other epidemic features also shape tree topology. Future work should evaluate performance under alternative assumptions about epidemic dynamics such as group transmission patterns [8], the effects of saturation [31], vaccination or new variants of concern [30], which would require developing additional distance metrics for PERMANOVA to capture such features. Our simulation framework focused on epidemics of 20–200 cases, reflecting the typical range for computational outbreak reconstruction, and our results show that PERMANOVA performs well once forests comprise 100 or more trees, corresponding to the typical effective sample size from Bayesian reconstruction models [12].

While alternative methods for comparing graph collections exist, they typically rely on abstract graph kernels not directly interpretable in our epidemiological context [32]. In contrast, our method employs a distance metric that is epidemiologically meaningful as it corresponds to the number of generations of infection separating each pair of cases. Furthermore, PERMANOVA can also be used for multifactorial analysis to quantify the relative contributions of the inference method, data type, and prior assumptions to the observed topological differences between epidemic forests. In addition to the application to epidemic

reconstruction that we have considered here, this work addresses a more general methodological gap across disciplines where relational structures are represented as graphs [33–38]. In practice, diverse data sources, modelling assumptions, and analytical methods typically produce not single solutions but ensembles of plausible alternatives, *i.e.* collections of graphs. Bayesian approaches excel at generating these collections through MCMC sampling but lack formal statistical tools for comparing the resulting posterior samples. One example of other such application area is phylogenetic tree reconstruction [38], where researchers encounter similar challenges that can lead to conflicting evolutionary hypotheses or taxonomic classifications. In information and network science, different network representations may likewise suggest distinctive social patterns or information flow dynamics.

In conclusion, our framework enables the comparison of collections of transmission trees, a special class of graph, by distinguishing meaningful structural variations from sampling and model uncertainty. We have demonstrated its utility to epidemic reconstruction, but this approach likely extends to other fields relying on graph-based representations. We encourage researchers to adapt and validate this framework to address domain-specific challenges in their respective fields, potentially developing additional metrics that capture the unique characteristics of their data structures.

## 6 Methods

We introduce a framework for comparing collections of transmission trees, termed *epidemic forests*. We present two approaches: the first based on a  $\chi^2$  test [22] on transmission pair frequencies, and the second using PERMANOVA, a method originally developed for ecological community analysis [23], on transmission tree distances. Both methods are described below and illustrated in Fig. 1. We use a simulation to compare the respective performances of the two approaches.

### 6.1 Epidemic Forests

Transmission trees represent the spread of a disease amongst infected individuals as directed acyclic graphs (DAGs) [1]. A transmission tree  $T = (V, E)$  consists of a set  $V = \{v_1, v_2, \dots, v_n\}$  containing  $n$  vertices (each representing an infected individual) and a set  $E = \{e_2, e_3, \dots, e_n\}$  of  $n - 1$  directed edges. Each edge represents an infector-infectee pair, denoted as  $e_j = (v_i, v_j)$ , with  $v_i, v_j \in V$  and  $v_i \neq v_j$ . This directed edge connects an infector  $v_i$  to its infectee  $v_j$ , formally encoding the ‘who infected whom’ relationship. All vertices have an in-degree of 1, except the root which represents the index case and has an in-degree of 0. In the absence of data to define meaningful edge weights, we assume all edges have a weight of 1.

We define an *epidemic forest* as a collection of transmission trees, each with the exact same set of vertices, but possibly different sets of edges. We consider two epidemic forests  $\mathcal{F}_A = (T_1^A, \dots, T_{m_A}^A)$  and  $\mathcal{F}_B = (T_1^B, \dots, T_{m_B}^B)$ , where the  $k^{\text{th}}$  tree in  $\mathcal{F}_A$  is defined as  $T_k^A = (V, E_k^A)$ . For simplicity, we assume that the two epidemic forests have the same size ( $m_A = m_B = m$ ), but the approaches described below can readily accommodate ( $m_A \neq m_B$ ). In practice, an epidemic forest may be obtained by sampling from a posterior distribution via Bayesian inference (*e.g.*, MCMC) or from a stochastic transmission model [12, 39].

### 6.2 $\chi^2$ test

The  $\chi^2$  test compares the absolute frequencies of infector-infectee pairs (*i.e.* edges) between two epidemic forests  $\mathcal{F}_A$  and  $\mathcal{F}_B$ . For each of the possible infector-infectee pair, we count their occurrences across all trees in a forest  $\mathcal{F}_X$  as:

$$c_{ij}^{\mathcal{F}_X} = \sum_{l=1}^m \mathbb{1}_{((v_i, v_j) \in E_l^X)} \quad (1)$$

where  $\mathbb{1}$  is the indicator function (yielding 1 if the pair appears in tree  $T_l^X$ , 0 otherwise).

The  $\chi^2$  statistic for comparing forests  $\mathcal{F}_A$  and  $\mathcal{F}_B$  is:

$$\chi^2 = \sum_{(i,j) \in \mathcal{P}} \frac{(c_{ij}^{\mathcal{F}_A} - c_{ij}^{\mathcal{F}_B})^2}{c_{ij}^{\mathcal{F}_A} + c_{ij}^{\mathcal{F}_B}} \quad (2)$$

where  $\mathcal{P} = \{(i, j) \mid i \neq j, c_{ij}^{\mathcal{F}_A} + c_{ij}^{\mathcal{F}_B} > 0\}$  includes only infector-infectee pairs observed in at least one forest. Under the null hypothesis that both forests stem from the same underlying frequency distribution of infector-infectee pairs,  $\chi^2$  follows a chi-square distribution with  $|\mathcal{P}| - 1$  degrees of freedom, where  $|\mathcal{P}|$  denotes the number of unique infector-infectee pairs observed. To accomodate small counts, the non-parametric Monte Carlo version of the chi-square test (999 replicates) was then used [22, 40]. This formulation assumes equal forest sizes ( $m_A = m_B = m$ ). Under the null hypothesis that both forests are sampled from the same distribution of infector-infectee pairs, the expected count for pair  $(i, j)$  in forest  $\mathcal{F}_A$  is  $E_{ij}^{\mathcal{F}_A} = \frac{c_{ij}^{\mathcal{F}_A} + c_{ij}^{\mathcal{F}_B}}{2}$ , and similarly for  $\mathcal{F}_B$ . Substituting these expected values into the classical chi-squared formula  $\frac{(O-E)^2}{E}$  and simplifying yields Equation 2.

### 6.3 PERMANOVA

PERMANOVA is a generic approach used to test group differences using pairwise distances between all observations of a sample and makes no model assumptions [23]. Here, we apply it to test whether distances between transmission trees differ when the trees belong to the same epidemic forest versus different forests.

#### 6.3.1 Distance between two transmission trees

The field of phylogenetics offers a range of established methods for comparing tree structures, providing several distance metrics for quantifying topological differences between pairs of phylogenies [17, 41–45]. These methods typically follow a two-step process: (i) convert trees into vectors of pairwise distances between all sampled taxa and (ii) compute Euclidean distances between these vectors.

A commonly used metric for the first step is the *patristic* distance [43], defined as the sum of branch lengths on the path separating two taxa, reflecting the evolutionary distance between them. Adapting this concept to transmission trees, we define the graph distance between cases (*i.e.* vertices)  $v_i$  and  $v_j$  as the sum of edge weights along their connecting path on the undirected graph. Since all edges here have a weight of 1, this distance directly corresponds to the number of transmission events between cases, carrying clear epidemiological meaning. An illustration of graph distances in a transmission tree is available in the supplementary material (Fig.S1).

We denote  $\pi(\cdot)$  the function mapping a transmission tree  $T$  of size  $n$  into a vector of  $\frac{n(n-1)}{2}$  graph distances:

$$\mathbf{d}_T = \pi(T) \quad (3)$$

where  $\mathbf{d}_T \in \mathbb{R}_+^{n(n-1)/2}$ .

The dissimilarity between two trees  $T_k$  and  $T_l$  is then quantified by the Euclidean distance between the respective vectors of graph distances, calculated

as the norm:

$$D(T_k, T_l) = \|\mathbf{d}_{T_k} - \mathbf{d}_{T_l}\| \quad (4)$$

This distance captures topological differences by evaluating how the relative positions of vertices, encoded as graph distances, diverge between the two trees. If  $T_k$  and  $T_l$  have identical edge sets, their graph distance matrices are equal, yielding  $D(T_k, T_l) = 0$ ; otherwise, discrepancies in path lengths increase the distance.

### 6.3.2 Outline of the method

Given two epidemic forests,  $\mathcal{F}_A$  and  $\mathcal{F}_B$ , each containing  $m$  transmission trees, we apply PERMANOVA to test whether tree topologies differ significantly between forests. Broadly, the method partitions pairwise distances between all trees into within-group ( $SS_W$ ) and between-group ( $SS_B$ ) components [23], based on pre-defined groups (here, the two forests). Statistical significance is assessed through permutation testing, where forest labels are randomly reassigned multiple times.

We define the combined epidemic forest as  $\mathcal{F}_{A \cup B} = \mathcal{F}_A \cup \mathcal{F}_B$ , containing all trees from  $\mathcal{F}_A$  and  $\mathcal{F}_B$ . The total sum of squares,  $SS_T$ , representing the overall variance across all trees in  $\mathcal{F}_{A \cup B}$ , is:

$$SS_T = \frac{1}{2m} \sum_{k=1}^{2m} \sum_{l=1}^{2m} D(T_k^{A \cup B}, T_l^{A \cup B})^2 \quad (5)$$

The double summation computes squared pairwise distances amongst the  $2m$  trees in  $\mathcal{F}_{A \cup B}$ , which decomposes to:

$$SS_T = \frac{1}{2m} \left( \sum_{k=1}^m \sum_{l=1}^m D(T_k^A, T_l^A)^2 + \sum_{k=1}^m \sum_{l=1}^m D(T_k^B, T_l^B)^2 + 2 \sum_{k=1}^m \sum_{l=1}^m D(T_k^A, T_l^B)^2 \right) \quad (6)$$

The within-group sum of squares  $SS_W$  measures the variance within forests:

$$SS_W = \frac{1}{m} \left( \sum_{k=1}^m \sum_{l=1}^m D(T_k^A, T_l^A)^2 + \sum_{k=1}^m \sum_{l=1}^m D(T_k^B, T_l^B)^2 \right) \quad (7)$$

where each term sums the squared distances among all pairs within each forest, normalised by  $m$ . The between-group sum of squares ( $SS_B$ ), capturing variability between the forests, is:

$$SS_B = SS_T - SS_W \quad (8)$$

The PERMANOVA test statistic [23] is:

$$F = \frac{SS_B}{SS_W / (2m - 2)} \quad (9)$$

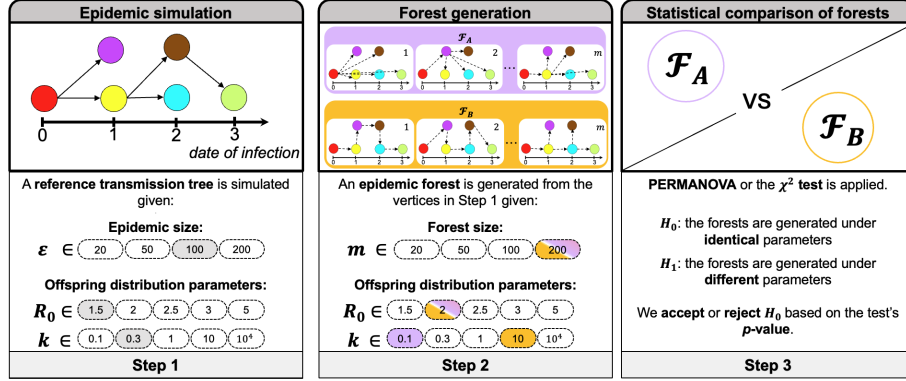


Figure 3: Simulation framework for assessing the performance of the  $\chi^2$  test and PERMANOVA.

Diagram illustrating the simulation study to assess the respective performances of  $\chi^2$  test and PERMANOVA for detecting differences between pairs of epidemic forests.

1. We simulate a *reference* transmission tree  $\mathcal{T}$  with  $\varepsilon$  infections from offspring distribution  $\text{NegBin}(R_0, k)$ . This process is repeated 100 times to account for the stochasticity of epidemic dynamics.
2. We generate reconstructed forests  $\mathcal{F}_A$  and  $\mathcal{F}_B$ , each containing  $m$  trees, by re-assigning infector-infectee relationships from  $\text{NegBin}(R'_{0,A}, k'_A)$  and  $\text{NegBin}(R'_{0,B}, k'_B)$ , conditional on  $\mathcal{T}$ 's dates of infection and case identifiers. In this example,  $\mathcal{F}_A = \text{NegBin}(R_0 = 2, k = 0.1)$  and  $\mathcal{F}_B = \text{NegBin}(R_0 = 2, k = 10)$ .
3. The  $\chi^2$  test and PERMANOVA are applied to test whether the two epidemic forests stem from the same generative process.

The reference distribution of  $F$  under the null hypothesis of no differences between groups is generated by a Monte Carlo procedure where forests labels are permuted a large number of times (*i.e.* 999 by default).  $p$ -values are calculated as the proportion of permuted  $F$ -values exceeding the observed  $F$  [23].

## 6.4 Simulation study

We conducted a simulation study to evaluate the performance of the  $\chi^2$  test and PERMANOVA in distinguishing between simulated epidemic forests drawn from distinct generative processes corresponding to different epidemic dynamics. The simulation framework is illustrated in Figure 3.



#### 6.4.1 Simulating epidemic forests

We generated epidemic forests through a three-stage process to systematically evaluate forest comparison methods across diverse transmission scenarios.

First, we defined the parameter space for the simulations:

- **Epidemic size:**  $\varepsilon \in \{20, 50, 100, 200\}$ . The number of infected individuals, corresponding to the number of vertices in the tree.
- **Basic Reproduction Number:**  $R_0 \in \{1.5, 2, 3\}$ . The mean number of secondary infections per case in a fully susceptible population, corresponding to the mean of the negative binomial offspring distribution.
- **Dispersion Parameter:**  $k \in \{0.1, 0.3, 0.5, 1, \infty\}$ . Controls heterogeneity in individual transmission, corresponding to the dispersion of the negative binomial offspring distribution. Lower values indicate greater overdispersion; as  $k \rightarrow \infty$ , the distribution converges to Poisson.

For each epidemic sizes  $\varepsilon$ , we defined offspring distributions  $\text{NegBin}(R_0, k)$  using all pairwise combinations of basic reproduction number  $R_0$  and dispersion parameter  $k$ .

Second, we generated *reference* transmission trees. For each parameter combination  $(\varepsilon, R_0, k)$  we simulated a reference transmission tree  $\mathcal{T}_{(R_0, k)}^\varepsilon$  using a stochastic branching process. Secondary infections per case were drawn from  $\text{NegBin}(R_0, k)$ , and generation times followed a gamma distribution with mean of 12 days and standard deviation of 6 days. We generated 100 replicate trees per parameter set to account for stochasticity. Simulations were initialised with 10,000 susceptible individuals, ran for a maximum of 365 days, and terminated upon reaching exactly  $\varepsilon$  infections, thereby excluding saturation effects. Within each reference tree  $\mathcal{T}_{(R_0, k)}^\varepsilon$ , infected individuals were assigned identifiers  $v \in \{1, \dots, \varepsilon\}$ , ordered by their dates of infection  $t_v$ .

Third, we constructed epidemic forests by re-assigning cases' ancestries. For each reference tree  $\mathcal{T}_{(R_0, k)}^\varepsilon$ , we generated forests  $\mathcal{F}_{\mathcal{T}_{(R_0, k)}^\varepsilon}(R'_0, k')$  by conditioning on the observed infection set  $\mathcal{I}_{\mathcal{T}_{(R_0, k)}^\varepsilon} = \{(v, t_v)\}_{v=1}^\varepsilon$  while resampling ancestries from  $\text{NegBin}(R'_0, k')$ . Each forest comprised of  $m = 200$  trees. This procedure yielded 15 distinct forests per reference tree (one for each offspring distribution pair  $(R'_0, k')$ ), including one forest matched the reference tree's generative process, where  $(R'_0, k') = (R_0, k)$ .

This procedure generated a total of 6,000 reference trees ( $|\varepsilon| \times |R_0| \times |k| \times \text{replicates} = 4 \times 3 \times 5 \times 100$ ), each generating 15 distinct forests ( $|R'_0| \times |k'|$ ), yielding 120 pairwise forest comparisons per reference tree ( $\binom{15}{2} + 15$ ), resulting in a total of 720,000 forest comparisons.

#### 6.4.2 Assessing statistical performance

For each of the 720,000 forest comparisons ( $\mathcal{F}_A$  vs.  $\mathcal{F}_B$ ), we performed the  $\chi^2$  test and PERMANOVA under 4 forest sizes  $m \in 20, 50, 100, 200$ , where  $m$  denotes the number of trees sampled from each forest. This resulted in a total of 5,760,000 tests performed. For each parameter combination, we measured:

- **Sensitivity:** The proportion of tests that correctly rejected the null hypothesis ( $H_0$ ) when comparing forests generated with different offspring distributions, *i.e.*,  $\mathcal{F}_T(R_0, k)$  vs.  $\mathcal{F}_T(R'_0, k')$  where  $(R_0, k) \neq (R'_0, k')$ .
- **Specificity:** The proportion of tests that correctly accepted  $H_0$  when comparing forests generated with identical offspring distributions, *i.e.*,  $(R_0, k) = (R'_0, k')$ .
- 

To quantify the factors influencing test sensitivity, we fit a logistic regression model to all comparisons where forests were generated under different parameter settings ( $H_1$ ;  $n = 5,040,000$ ). The binary outcome was whether the test correctly rejected the null hypothesis ( $H_0$ ). We compared four nested models using the Akaike Information Criterion and selected the model with the lowest value. The final model included main effects for statistical method (PERMANOVA or  $\chi^2$ ), forest size ( $m$ ), epidemic size ( $\varepsilon$ ), and parameter differences between forests ( $\Delta R_0$  and  $\Delta k$ ). It also included all two way and three way interaction terms involving the method:

$$\begin{aligned} \text{logit}(P(\text{reject } H_0)) = & \beta_0 + \beta_{\text{method}} + \beta_m + \beta_\varepsilon + \beta_{\Delta R_0} + \beta_{\Delta k} \\ & + \beta_{\text{method}:\varepsilon} + \beta_{\text{method}:\Delta R_0} + \beta_{\text{method}:\Delta k} \\ & + \beta_{\Delta R_0:\Delta k} + \beta_{\text{method}:\Delta R_0:\Delta k} + e \end{aligned} \quad (10)$$

Where ‘.’ represent the interaction term and  $e$  is the normally distributed residuals. Results are reported as odds ratios using the  $\chi^2$  test, the smallest forest size ( $m = 20$ ), the smallest epidemic size ( $\varepsilon = 20$ ), and no difference in dispersion ( $\Delta k = 0$ ) as reference categories. The model achieved a pseudo  $R^2$  of 0.58 [24].

Both methods achieved near-perfect specificity ( $> 97\%$ ) across all conditions, precluding regression analysis.

## 7 Author contributions

- Conceptualisation: CG, AC, TJ
- Methodology: CG, AC, TJ
- Software: CG
- Validation: CG, AC, TJ
- Visualization: CG
- Writing – original draft: CG
- Writing – review & editing: CG, AC, TJ
- Supervision: AC, PJW, TJ
- Funding acquisition: PJW

## 8 Funding and competing interests

This work was funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics, which was a partnership between Imperial College London, London School of Hygiene and Tropical Medicine, and UKHSA (grant code NIHR200908). All authors declare that they have no competing interests.

## 9 Data and materials availability

Simulations, analyses and visualisations were performed using the R software version 4.4.0 (<https://www.R-project.org/>) [26]. Our framework has been implemented in a free, open-source, R package *mixtree*, which is available on CRAN [27]. This study is fully reproducible using code available on GitHub: [https://github.com/CyGei/mixtree\\_analysis](https://github.com/CyGei/mixtree_analysis) (archived on Zenodo:10.5281/zenodo.17704758 [46]). The resulting data is stored on a Zenodo archive: 10.5281/zenodo.17704455 [47].

## References

- [1] T. Jombart et al. “Reconstructing disease outbreaks from genetic data: a graph approach”. en. In: *Heredity* 106.2 (Feb. 2011), pp. 383–390.
- [2] Liang Wang et al. “Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase”. In: *Nature communications* 11.1 (2020), p. 5006.
- [3] Thomas R. Frieden and Christopher T. Lee. “Identifying and Interrupting Superspreading Events—Implications for Control of Severe Acute Respiratory Syndrome Coronavirus 2”. en. In: *Emerging Infectious Diseases* 26.6 (June 2020), p. 1059.
- [4] J. O. Lloyd-Smith et al. “Superspreading and the effect of individual variation on disease emergence”. In: *Nature* 438.7066 (Nov. 2005), pp. 355–359.
- [5] Mohamed Abbas et al. “Reconstruction of transmission chains of SARS-CoV-2 amidst multiple outbreaks in a geriatric acute-care hospital: a combined retrospective epidemiological and genomic study”. In: *eLife* 11 (July 2022). Ed. by Joshua T Schiffer et al., e76854.
- [6] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. “Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data”. In: *Molecular Biology and Evolution* 31.7 (July 2014), pp. 1869–1879.
- [7] Xavier Didelot et al. “Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks”. In: *Molecular Biology and Evolution* 34.4 (Apr. 2017), pp. 997–1007.
- [8] Cyril Geismar et al. “Sorting out assortativity: When can we assess the contributions of different population groups to epidemic transmission?” en. In: *PLOS ONE* 19.12 (Dec. 2024), e0313037.
- [9] Mohamed Abbas et al. “Explosive nosocomial outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak reconstruction”. In: *Journal of Hospital Infection* 117 (2021), pp. 124–134.
- [10] Cécile Kremer et al. “Reconstruction of SARS-CoV-2 outbreaks in a primary school using epidemiological and genomic data”. In: *Epidemics* 44 (Sept. 2023), p. 100701.
- [11] Finlay Campbell et al. “When are pathogen genome sequences informative of transmission events?” en. In: *PLOS Pathogens* 14.2 (Feb. 2018), e1006885.
- [12] Hélène Duault, Benoit Durand, and Laetitia Canini. “Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review”. In: *Pathogens* 11.2 (2022), p. 252.
- [13] Nicola De Maio, Chieh-Hsi Wu, and Daniel J Wilson. “SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent”. In: *PLoS computational biology* 12.9 (2016), e1005130.

- [14] Andrew Gelman and Donald B. Rubin. “Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4 (Nov. 1992), pp. 457–472.
- [15] Ben Lambert. “A student’s guide to Bayesian statistics”. In: (2018).
- [16] Joseph Felsenstein. “CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP”. In: *Evolution* 39.4 (July 1985), pp. 783–791.
- [17] Thibaut Jombart et al. “treespace: Statistical exploration of landscapes of phylogenetic trees”. en. In: *Molecular Ecology Resources* 17.6 (2017), pp. 1385–1392.
- [18] Thibaut Jombart et al. “Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data”. en. In: *PLOS Computational Biology* 10.1 (Jan. 2014), e1003457.
- [19] Don Klinkenberg et al. “Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks”. In: *PLoS computational biology* 13.5 (2017), e1005495.
- [20] Alan Gibbons. *Algorithmic Graph Theory*. en. June 1985.
- [21] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. “Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set”. en. In: *PLOS Computational Biology* 11.12 (Dec. 2015), e1004613.
- [22] Karl Pearson. “X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.
- [23] Marti J. Anderson. “A new method for non-parametric multivariate analysis of variance”. en. In: *Austral Ecology* 26.1 (2001), pp. 32–46.
- [24] Tue Tjur. “Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination”. In: *The American Statistician* 63.4 (Nov. 2009), pp. 366–372.
- [25] Jari Oksanen et al. *vegan: Community Ecology Package*. Jan. 2025.
- [26] R R Core Team. “R: A language and environment for statistical computing”. In: (2025).
- [27] Cyril Geismar. *mixtree: A Statistical Framework for Comparing Sets of Trees*. Mar. 2025.
- [28] Finlay Campbell et al. “Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data”. en. In: *PLOS Computational Biology* 15.3 (Mar. 2019), e1006930.
- [29] Jørgen Bang-Jensen and Gregory Z. Gutin. “Connectivity of Digraphs”. en. In: *Digraphs: Theory, Algorithms and Applications*. Ed. by Jørgen Bang-Jensen and Gregory Z. Gutin. 2009, pp. 191–226.

- [30] Cyril Geismar et al. “Bayesian reconstruction of SARS-CoV-2 transmissions highlights substantial proportion of negative serial intervals”. In: *Epidemics* 44 (2023), p. 100713.
- [31] Anne Cori et al. “A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics”. In: *American Journal of Epidemiology* 178.9 (Nov. 2013), pp. 1505–1512.
- [32] Ragnar L. Gudmundarson and Gareth W. Peters. “GTST: A Python Package for Graph Two-Sample Testing | Journal of Open Research Software”. en. In: (Jan. 2024).
- [33] Jonathan L. Gross, Jay Yellen, and Mark Anderson. *Graph Theory and Its Applications*. 3rd ed. Nov. 2018.
- [34] Peter Pin-Shan Chen. “The entity-relationship model—toward a unified view of data”. en. In: *ACM Transactions on Database Systems* 1.1 (Mar. 1976), pp. 9–36.
- [35] Daniel R. Zerbino and Ewan Birney. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome Research* 18.5 (May 2008), pp. 821–829.
- [36] Alexandru T. Balaban. “Applications of graph theory in chemistry”. en. In: *Journal of Chemical Information and Computer Sciences* 25.3 (Aug. 1985), pp. 334–343.
- [37] Mark S. Granovetter. “The Strength of Weak Ties”. In: *American Journal of Sociology* 78.6 (1973), pp. 1360–1380.
- [38] Alexei J. Drummond and Andrew Rambaut. “BEAST: Bayesian evolutionary analysis by sampling trees”. In: *BMC Evolutionary Biology* 7.1 (Nov. 2007), p. 214.
- [39] H. W. Watson and Francis Galton. “On the Probability of the Extinction of Families”. In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 4 (1875), pp. 138–144.
- [40] Drake R. Bradley and Steven Cutcomb. “Monte Carlo simulations and the chi-square test of independence”. en. In: *Behavior Research Methods & Instrumentation* 9.2 (Mar. 1977), pp. 193–201.
- [41] Sandrine Pavoine et al. “Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities”. In: *Theoretical Population Biology* 73.1 (Feb. 2008), pp. 79–91.
- [42] D. F. Robinson and L. R. Foulds. “Comparison of weighted labelled trees”. en. In: *Combinatorial Mathematics VI*. Ed. by A. F. Horadam and W. D. Wallis. 1979, pp. 119–126.
- [43] Mike A. Steel and David Penny. “Distributions of Tree Comparison Metrics—Some New Results”. In: *Systematic Biology* 42.2 (June 1993), pp. 126–141.
- [44] C. Colijn and G. Plazzotta. “A Metric on Phylogenetic Tree Shapes”. In: *Systematic Biology* 67.1 (Jan. 2018), pp. 113–126.

- [45] Michelle Kendall et al. “Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees”. In: *Statistical Science* 33.1 (2018), pp. 70–85.
- [46] CyGei. *CyGei/mixtree-analysis: mixtree-analysis*. Nov. 2025.
- [47] Cyril Geismar. *mixtree-analysis-data*. eng. Nov. 2025.
- [48] Claude E Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

## Supplementary Information

The following presents the supplementary materials for the paper entitled: ‘*A statistical framework for comparing epidemic forests*’. The first part illustrates the computation of graph distances in transmission trees, the second details the methodology for simulating epidemic forests, and the third presents additional results.

### S1 Graph distances

We define the graph distance between cases as the number of undirected edges along their connecting path, corresponding to the number of transmission events between the considered cases. The diagram below illustrates this concept using a transmission tree (panel A) with five cases. Panel B shows the corresponding matrix of graph distances between all pairs of cases. Distances between case 5 and all other cases are highlighted in colour, corresponding to the coloured paths in panel A. This process is conducted in step 1 of the right panel in Figure 1 of the main text.



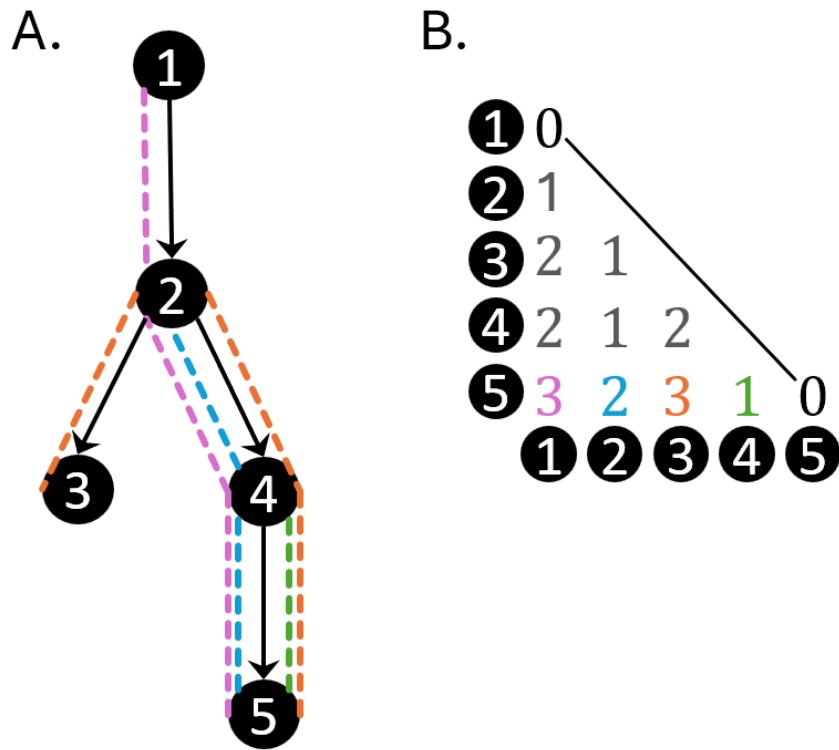


Figure S1: Calculation of graph distances in a transmission tree.

A. A transmission tree with 5 cases. The coloured dashed lines show the unique paths connecting case 5 to all other cases. B. The matrix representation of graph distances between all pairs of cases. The coloured numbers correspond to the number of transmission events between case 5 and all other cases, matching the coloured paths shown in A. For illustration, we only represent the lower triangle but the matrix is symmetric.

## S2 Simulation Framework

Our simulation study followed a two-step process designed to evaluate how effectively the proposed statistical methods could distinguish between epidemic forests derived from distinct generative processes. First, we simulated *reference* transmission trees using a branching process model with varying offspring distributions. Second, we generated epidemic forests from these reference trees by re-assigning infector-infectee relationships from a new offspring distribution, conditional on the reference’s tree case identifiers and dates of infection, producing collections of plausible transmission trees that reflect the underlying dynamics of each scenario. This approach provided a controlled environment with known ground truth against which we could systematically assess the discriminatory power of our statistical tests across different epidemic sizes, forest sizes, and offspring distribution parameters.

### S2.1 Branching process model

Our simulations employ the discrete-time stochastic branching process implemented in *simulacr* (<https://github.com/CyGei/simulacr>). Each simulation was initiated with a single infected individual (the index case), from whom infections propagated across successive generations based on specified offspring and generation time distributions.

*simulacr* tracks the propagation of infections through successive generations by computing the force of infection (FOI) at each time step  $[t, t + 1)$ . For each infected individual  $i$ , we define  $t_i$  as their infection time and  $R_i$  as their case reproduction number—the expected number of secondary cases they generate in a fully susceptible population. The generation time—the interval between the infection of a primary case and the infection of its secondary cases—follows a probability mass function  $g(t)$ , where  $g(t) = 0$  for  $t \leq 0$ . At each time step, the model sums the infectious contribution from all active cases to determine the overall force of infection, then probabilistically generates new infections from the susceptible population ( $S(t)$ ) and assigns their respective infectors according to their relative contribution to transmission.

The FOI generated by case  $i$  at time  $t$  is defined as:

$$\lambda_i(t) = R_i g(t - t_i) \quad (11)$$

The total FOI at time  $t$  arising from all infectious individuals at this time (denoted by the set  $I(t)$ ) is given by:

$$\Lambda(t) = \sum_{i \in I(t)} \lambda_i(t) \quad (12)$$

Each susceptible individual  $j$  (with  $j \in S(t)$ ) faces an infection probability during the interval  $[t, t + 1)$  of:

$$p_j = 1 - e^{-\Lambda(t)} \quad (13)$$

Once a susceptible individual becomes infected at time  $t + 1$ , a specific infector is drawn from a multinomial distribution where the probability that case  $i \in I(t)$  is the infector is defined as:

$$P(i \mid \text{infection at } t + 1) = \frac{\lambda_i(t)}{\Lambda(t)} \quad (14)$$

To achieve exactly  $\varepsilon$  cases per outbreak, we initialised simulations with 10,000 susceptible individuals and terminated them upon reaching exactly  $\varepsilon$  infections, thereby excluding saturation effects. This truncation scheme led to right-censoring, where cases infected within one generation time of the truncation date likely had not realised their expected number of secondary infections  $R_i$ .

For all simulations, we modelled the generation time using a discretised probability mass function (PMF) with a mean of 12 days and a standard deviation of 6 days to enable considerable entropy across reconstructed forests (supplementary Fig. S5).

Each simulated transmission tree returns the case identifiers of infected individuals and their infection dates which will be used for forest generation.

## S2.2 Forest generation

For each simulated reference transmission tree  $\mathcal{T}_{(R_0, k)}^\varepsilon$ , we generated epidemic forests  $\mathcal{F}_{\mathcal{T}_{(R_0, k)}^\varepsilon}$  by reassigning infector-infectee relationships given a new offspring distribution  $\text{NegBin}(R'_0, k')$  while conditioning on the reference tree's set of infection  $\mathcal{I}_{\mathcal{T}_{(R_0, k)}^\varepsilon} = \{(v, t_v)\}_{v=1}^\varepsilon$ , where  $v$  denotes case identifiers and  $t_v$  their infection times. Each forest comprised of  $m = 200$  trees.

For each tree in the forest, we followed a three-step process:

- **Offspring sampling:** we drew individual reproduction numbers  $R_i \sim \text{NegBin}(R'_0, k')$  for all  $\varepsilon$  cases. To ensure that the epidemic sets off, we constrained the index (root) case to have  $R_1 \geq 1$  by sampling from a truncated negative binomial distribution when  $R_1 < 1$ .
- **Force of infection (FOI) calculation:** For each case  $i$ , with infection time  $t_i$  and reproduction number  $R_i$ , we computed the FOI exerted by  $i$  at time  $t$  as:

$$\lambda_i(t) = R_i g(t - t_i)$$

where  $g()$  is the generation time probability mass function.

- **Ancestry assignment:** cases were assigned infectors sequentially in order of their infection times. For each case  $j$ , infected at time  $t_j$ , we identified the set of ancestors  $\mathcal{A}(t_j) = \{i : t_i < t_j\}$  and selected the infector by sampling from the multinomial distribution with probability relative to the contribution of  $i$  on the whole FOI at time  $t_j$ :

$$P(i \rightarrow j) = \frac{\lambda_i(t_j)}{\sum_{i' \in \mathcal{A}(t_j)} \lambda_{i'}(t_j)}$$

This procedure ensures that each generated tree  $\mathcal{T}_k = (V, E_k)$  in forest  $F_{\mathcal{T}}(R'_0, k')$  shares the same vertex set  $V$  and infections times  $t_v$  as the reference tree  $\mathcal{T}$ , but with edges resampled according to the FOI generated under the new offspring distribution  $\text{NegBin}(R'_0, k')$ . An epidemic forest is thus a collection of transmission trees that share the same set of infected individuals but different ancestral relationships, all consistent with a given offspring distribution.

### S3 Additional Results

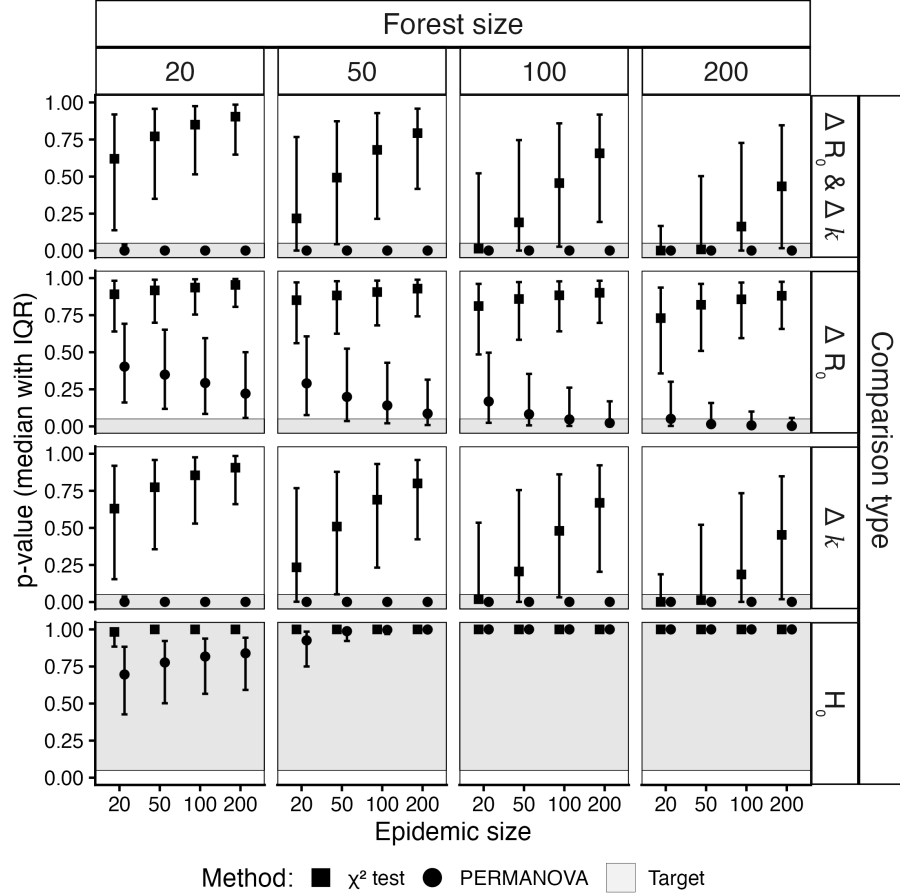


Figure S2: Performance of  $\chi^2$  test and PERMANOVA for distinguishing epidemic forests.

Median p-values and interquartile ranges for the  $\chi^2$  test (squares) and PERMANOVA (circles) across epidemic sizes (x-axis), forest sizes (columns), and parameter conditions (rows). Grey shading indicates desired p-value ranges: below  $\alpha = 0.05$  when forests differ in at least one parameter (rows 1–3, reject  $H_0$ ) and above  $\alpha = 0.05$  when forests share identical parameters (row 4, accept  $H_0$ ).

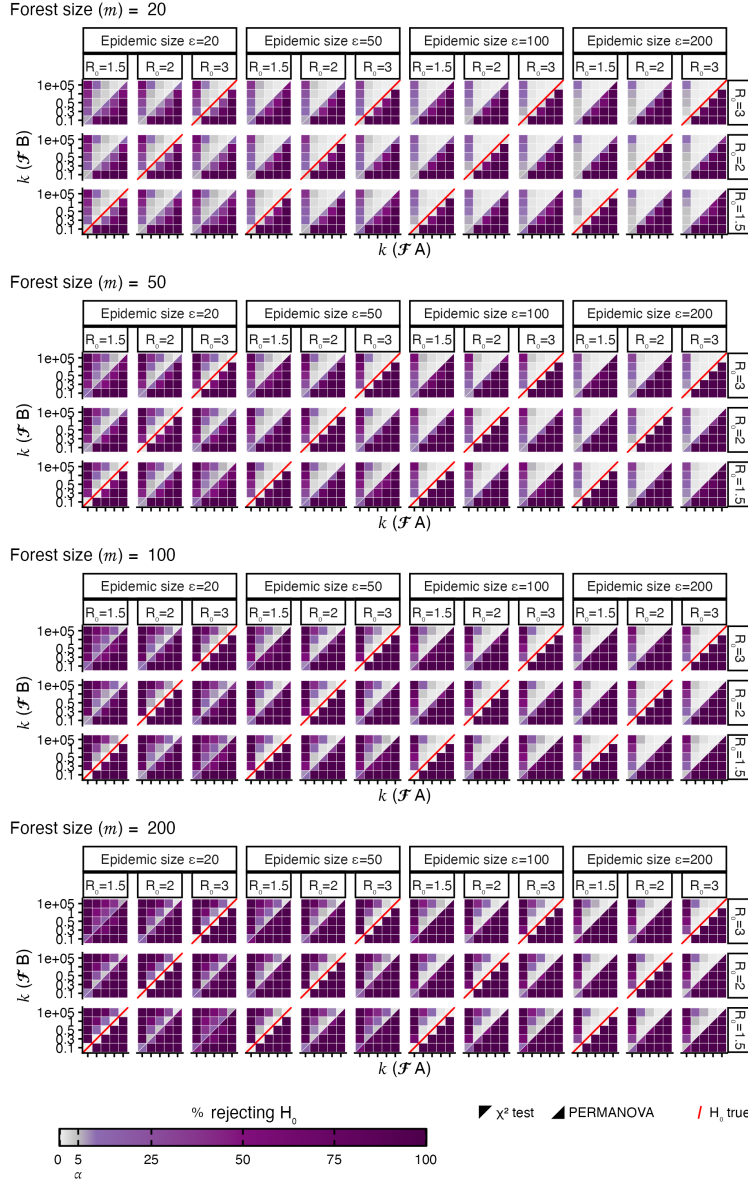


Figure S3: Performance of  $\chi^2$  and PERMANOVA in distinguishing epidemic forests

Each panel shows the proportion of tests rejecting the null hypothesis ( $p < 0.05$ ) when comparing epidemic forest  $\mathcal{F}_A$  and  $\mathcal{F}_B$ . The upper triangle shows the  $\chi^2$  test results; the lower triangle shows PERMANOVA results. Outer columns refer to epidemic size ( $\epsilon$ ), common to both forest. Forests can differ in their offspring distribution parameter:  $R_0$  (inner columns) and  $k$  (x and y axes; x-axis labels omitted for clarity, values identical to the y-axis). Red diagonal lines indicate comparisons where both forests share identical parameters ( $H_0$  true; low rejection rates indicate good specificity). The other cells compare forests with different parameters (high rejection rates indicate good sensitivity). Both methods maintain excellent specificity (diagonal), but PERMANOVA demonstrates superior sensitivity.

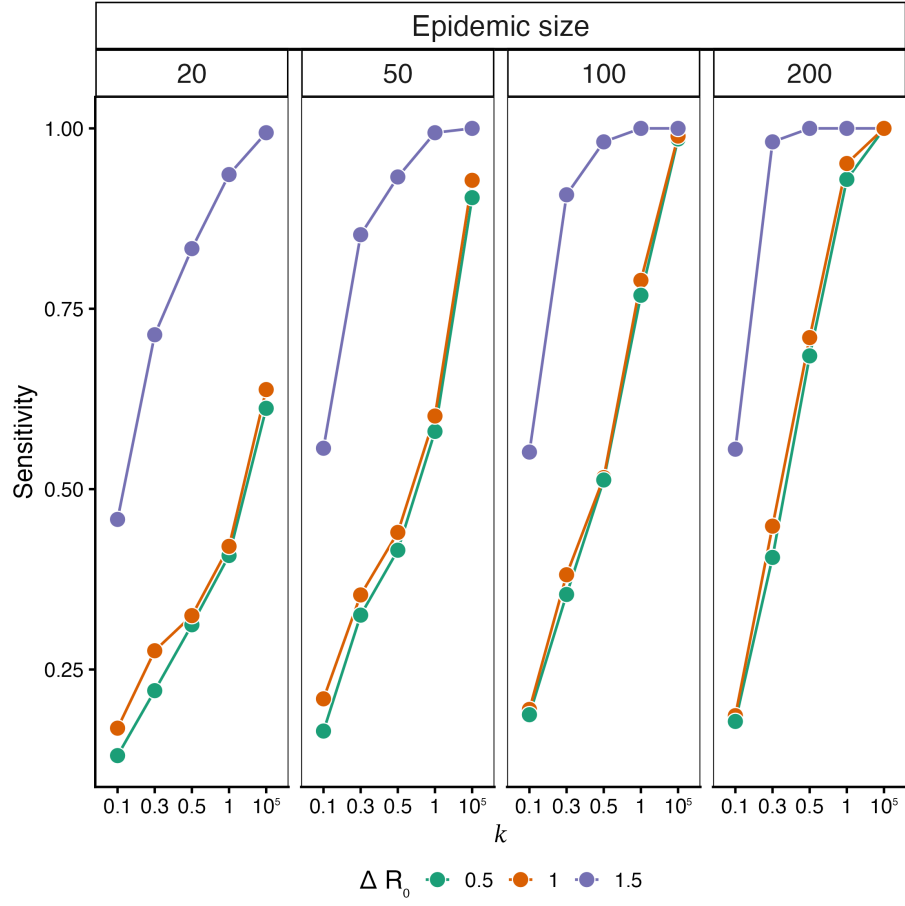


Figure S4: Sensitivity of PERMANOVA when comparing epidemic forests that differ only in  $R_0$ .

Sensitivity (y-axis) is the proportion of tests correctly rejecting the null hypothesis when comparing forests of 200 trees generated with different  $R_0$  ( $\Delta R_0$ , colour) but identical  $k$  (x-axis). Columns correspond to epidemic size.

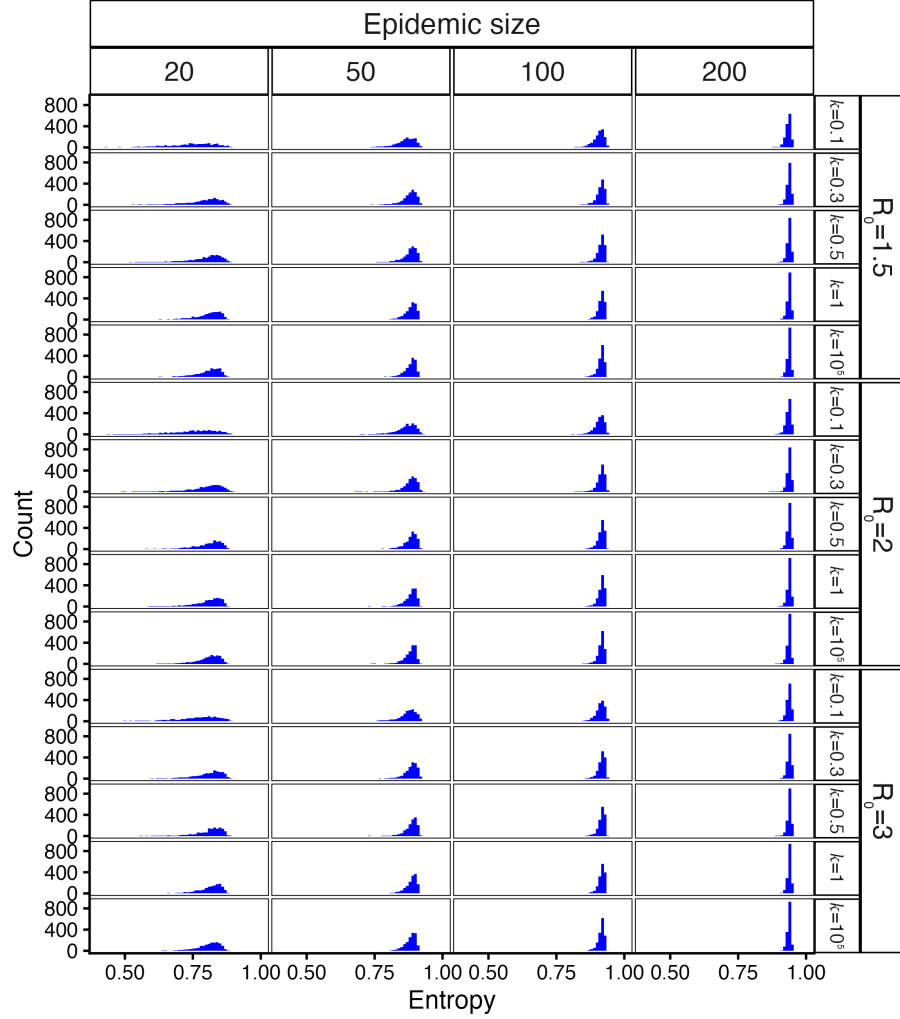


Figure S5: Variation in infector-infectee relationships across epidemic forests

Histogram of mean scaled entropy across epidemic forests ( $m = 200$  trees per forest), stratified by simulation parameters. For each infectee  $j$ , the scaled entropy  $H_j$  (x axis) quantifies variation in their assigned infector across all trees in a forest, computed using the normalised Shannon entropy formula [48]:  $H_j = \frac{-\sum_{i=1}^{K_j} p_{ij} \log(p_{ij})}{\log(K_j)}$ , where  $p_{ij}$  is the proportion of trees in which individual  $i$  infects  $j$ , and  $K_j$  is the number of distinct infectors of  $j$  observed across the forest. Values range from 0 (identical infector in all trees) to 1 (all possible infectors equally frequent). The mean scaled entropy ( $\bar{H}$ ) is obtained by averaging  $H_j$  over all cases. Columns refer to epidemic sizes ( $\varepsilon$ ), rows refer to the mean reproduction number  $R_0$  and dispersion parameter  $k$  of the negative binomial offspring distribution. Average entropy for our simulations is 77% and increases with epidemic size, due to greater variation in infector assignment (TableS1).



The mean scaled entropy  $\bar{H}$  for each forest was modelled as a linear function of epidemic size, reproduction number, and dispersion:

$$\bar{H}^* = \beta_0 + \beta_\varepsilon + \beta_{R_0} + \beta_k + \epsilon \quad (15)$$

where  $\beta_0$  is the intercept, each  $\beta$  term represents the categorical effect of epidemic size ( $\varepsilon$ ), reproduction number ( $R_0$ ), and dispersion parameter ( $k$ ) respectively, and  $\epsilon$  is the residual error. The model explained 68.2% of the variance ( $R^2 = 0.682$ ) in mean scaled entropy across 90,000 simulated forests, with coefficient estimates shown in Table S1.

Table S1: Linear regression results for the mean scaled entropy ( $\bar{H}$ ) model (Eq. 15)

Predictor		Estimate	p-value
Intercept		0.771	<0.001
<b>Epidemic size (<math>\varepsilon</math>)</b>	50	0.087	<0.001
	100	0.123	<0.001
	200	0.147	<0.001
<b>Reproduction number (<math>R</math>)</b>	2	0.003	<0.001
	3	0.006	<0.001
	0.3	0.018	<0.001
<b>Dispersion (<math>k</math>)</b>	0.5	0.020	<0.001
	1	0.021	<0.001
	Poisson	0.021	<0.001

*Note:* Model fit  $R^2 = 0.682$ . All p-values are < 0.001 due to the large simulation sample size ( $n = 90,000$ ). Coefficient estimates indicate the expected change in mean scaled entropy relative to the reference category ( $\varepsilon = 20$ ,  $R = 1.5$ ,  $k = 0.1$ ).

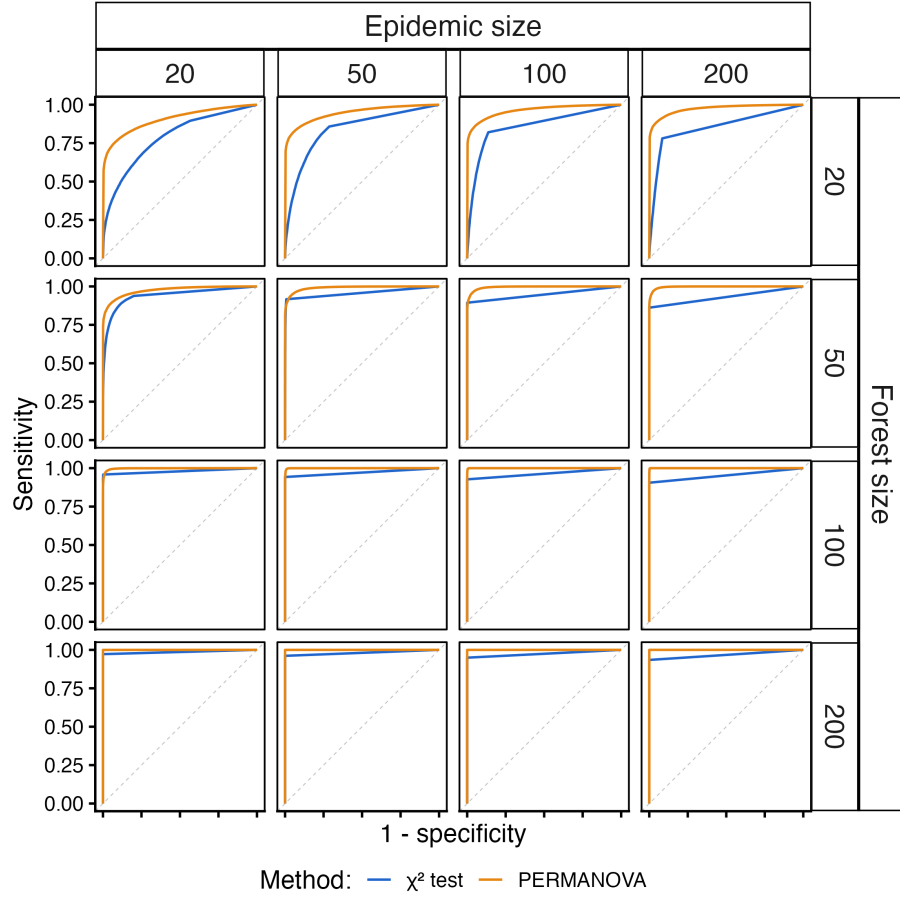


Figure S6: ROC curves for the  $\chi^2$  test and PERMANOVA

Each panel shows the receiver operating characteristic (ROC) curves plotting true positive rate (sensitivity) against false positive rate (1-specificity) for the  $\chi^2$  test (blue) and PERMANOVA (orange) across all simulations for all possible significance thresholds ( $0 \leq \alpha \leq 1$ ). Panels are arranged by epidemic size (columns: 20–200 cases) and forest size (rows: 20–200 trees), x-axis tick labels are omitted for clarity, as both axes share the same scale.

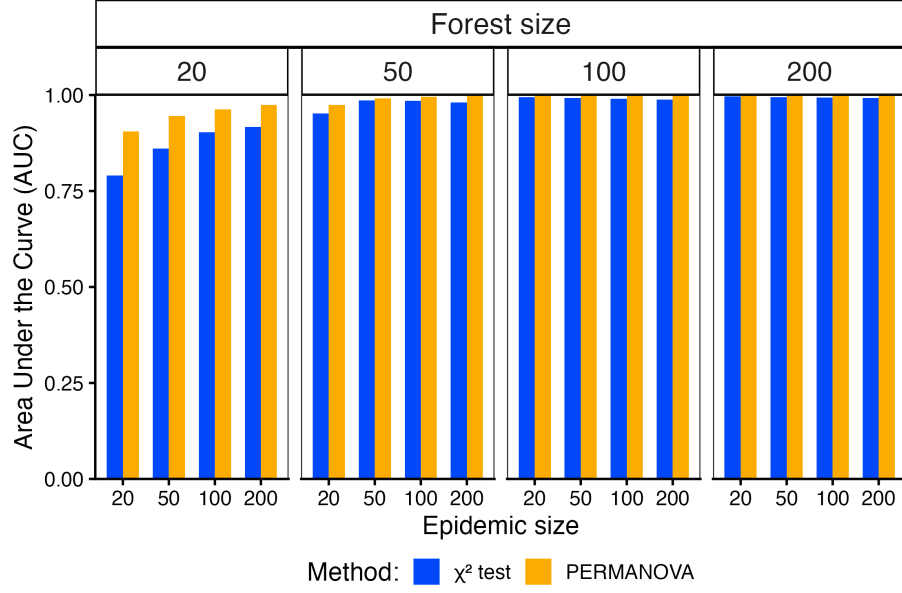


Figure S7: Area under curve (AUC) for  $\chi^2$  test and PERMANOVA

This figure shows the AUC derived from the Receiver Operating Characteristic (ROC) curves (supplementary Fig. S6) of the two tests evaluated in our simulations. The y-axis displays the AUC value, with higher values corresponding to better performances. An AUC of 1 corresponds to a test with perfect sensitivity and specificity. The x-axis displays the epidemic size *i.e.* the number of cases in the simulated epidemics, while panels refer to the forest size *i.e.* the number of trees in each forest.

method	min	lower quartile	mean	median	upper quartile	max
PERMANOVA	5.16	5.48	5.62	5.59	5.74	6.33
$\chi^2$ test	0.52	0.53	0.53	0.53	0.54	0.58

Table S2: Benchmark results of execution times in seconds for the  $\chi^2$  test and PERMANOVA, comparing two epidemic forests with 100 trees and 100 vertices each. Both tests used 999 permutations (PERMANOVA) / Monte Carlo replicates ( $\chi^2$  test) without parallelisation and were replicated 100 times per method.

For the  $\chi^2$  test, we compute the frequency of each infector-infectee pair across all trees between forests. In the worst-case scenario, where every possible infector-infectee pair (*i.e.*  $n(n-1)$  pairs) appears at least once in either forest, the computational time for the  $\chi^2$  test increases with the number of trees in each forest ( $m$ ) and the square of the number of cases ( $n^2$ ), since it considers all infector-infectee pairs for every tree (See Methods, Fig.1). Therefore the overall computational time will increase as a function of  $mn^2$ .

On the other hand, PERMANOVA involves a two-step process. First, it computes pairwise distances between all vertices within each tree (Fig. S1), which scales as a function of  $n^2$  for a given tree. Second, it calculates pairwise distances between all trees (Fig.1), which scales as a function of  $m^2$ . Therefore the overall computational time will increase as a function of  $m^2n^2$ .