Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

Fabio Pardi, Celine Scornavacca

Reconstructible phylogenetic networks: do not distinguish the indistinguishable

Fabio Pardi[1,3*], Celine Scornavacca[2,3]

1 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM, UMR 5506) CNRS, Université de Montpellier, France, **2** Institut des Sciences de l'Evolution de Montpellier (ISE-M, UMR 5554) CNRS, IRD, Université de Montpellier, France, **3** Institut de Biologie Computationnelle, Montpellier, France

* fabio.pardi@lirmm.fr

Abstract

Phylogenetic networks represent the evolution of organisms that have undergone reticulate events, such as recombination, hybrid speciation or lateral gene transfer. An important way to interpret a phylogenetic network is in terms of the trees it displays, which represent all the possible histories of the characters carried by the organisms in the network. Interestingly, however, different networks may display exactly the same set of trees, an observation that poses a problem for network reconstruction: from the perspective of many inference methods such networks are indistinguishable. This is true for all methods that evaluate a phylogenetic network solely on the basis of how well the displayed trees fit the available data, including all methods based on input data consisting of clades, triples, quartets, or trees with any number of taxa, and also sequence-based approaches such as popular formulations of maximum parsimony and maximum likelihood for networks. This identifiability problem is partially solved by accounting for branch lengths, although this merely reduces the frequency of the problem. Here we propose that network inference methods should only attempt to reconstruct what they can uniquely identify. To this end, we introduce a novel definition of what constitutes a uniquely reconstructible network. For any given set of indistinguishable networks, we define a canonical network that, under mild assumptions, is unique and thus representative of the entire set. Given data that underwent reticulate evolution, only the canonical form of the underlying phylogenetic network can be uniquely reconstructed. While on the methodological side this will imply a drastic reduction of the solution space in network inference, for the study of reticulate evolution this is a fundamental limitation that will require an important change of perspective when interpreting phylogenetic networks.

Author Summary

We consider here an elementary question for the inference of phylogenetic networks: what networks can be reconstructed. Indeed, whereas in theory it is always possible to reconstruct a phylogenetic tree, given sufficient data for this task, the same does not hold for phylogenetic networks: most notably, the relative order of consecutive reticulate events cannot be determined by standard network inference methods. This problem has been described before, but no solutions to deal with it have been put forward. Here we propose limiting the space of reconstructible phylogenetic networks to what we call "canonical networks". We formally prove that each network has a (usually unique) canonical form – where a number of nodes and branches are merged – representing all that can be uniquely reconstructed about the original network. Once a canonical network $\hat{N}$ is inferred, it must be kept in mind that – even with perfect and unlimited data – the true phylogenetic network is just one of the potentially many networks having $\hat{N}$ as canonical form. This is an important difference to what biologists are used to for phylogenetic trees, where in principle it is always possible to resolve uncertainties, given enough data.

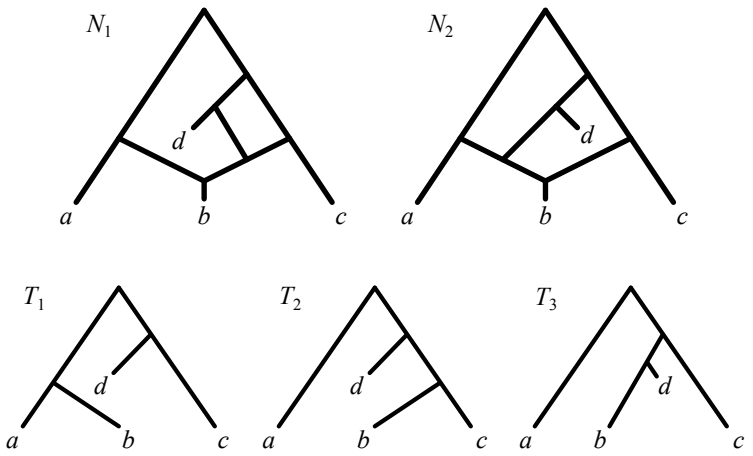


Fig 1. Indistinguishable network topologies. The network topologies N_1 and N_2 are indistinguishable to most current approaches for network reconstruction, as they display the same tree topologies T_1, T_2 and T_3.

Introduction

Explicit [1] or *evolutionary* [2,3] phylogenetic networks are used to represent the evolution of organisms or genes that may inherit genetic material from more than one source. This may be caused by events such as hybrid speciation (e.g. in plants and animals [4,5]), horizontal gene transfer (e.g. in bacteria [6,7]), viral reassortment [8], or recombination (e.g. in viruses [9,10] or in the genomes of sexually reproducing species [11–13]). They are called "explicit" to distinguish them from "implicit" [14], "abstract" [1] or "data-display" [3] phylogenetic networks, which are used to display collections of alternative evolutionary hypotheses supported by conflicting signals in the data. In explicit networks, multiple-inheritance events are represented as *reticulations*, that is, nodes where two or more lineages converge to give rise to a new lineage, whose genetic material is a combination of that of its direct ancestors.

Explicit networks can be interpreted in terms of classic, tree-like evolution: if we focus on a single, indivisible and thus non-recombining inherited character (for example a single site in a DNA sequence), its history is still best described by a tree. This observation gives rise to the notion of *trees displayed by a network*, which are all the possible single-character histories implied by a phylogenetic network. (See, e.g., Fig. 1, where T_1, T_2 and T_3 are the trees displayed by networks N_1 and N_2. Formal definitions are in the Results section.)

Several works in the last few years have focused on the methodology for phylogenetic network inference, and data-display networks in particular have begun to make a real impact on the everyday practice of biologists (e.g., [15–17]). There remains, however, a strong demand for automatic reconstruction of networks that not only display conflicting signals in the data, but also seek to explain these signals with explicit inferences of past reticulation events (see, e.g., [18–20]). This is evidenced, for example, by the abundance of manually reconstructed networks in the literature [8,21–27]. As a result of this demand, the inference of explicit networks is now a rapidly growing field of research [1].

Some paradigms in the proposed methodology are beginning to emerge. Not surprisingly, the notion of trees displayed by a phylogenetic network plays a central role: the general idea is to evaluate the fit of a network N with the data *indirectly* – on the basis of how well the trees displayed by N explain the data. In the following, we describe how this applies to the two main approaches for explicit network reconstruction: consistency-based approaches (see [28] for a survey) – seeking a network consistent with a number of prior evolutionary inferences (typically trees or groupings of taxa) – and sequence-based approaches, such as

standard formulations of maximum parsimony and maximum likelihood for networks [2, 29–33].

Although evaluating a network via the trees it displays is evolutionarily meaningful, it has a problematic consequence: from the perspective of these reconstruction methods, all networks displaying the same set of trees are "indistinguishable", as the function that these methods seek to optimize will always assign the same score to all networks displaying the same set of trees, regardless of the input data. In other words, the central parameter of phylogenetic network inference, the network itself, is in some cases not identifiable.

An Identifiability Problem

As an example, consider again networks N_1 and N_2 in Fig. 1, which display the same trees $\mathcal{T}(N) = \{T_1, T_2, T_3\}$. (In the following, $\mathcal{T}(N)$ denotes the set of trees displayed by N.) By displaying the same trees, these networks display the same clades, the same triples, the same quartets (triples and quartets are rooted subtrees with 3 leaves and unrooted subtrees with 4 leaves, respectively) and in general the same subtrees with an arbitrary number of leaves. Therefore, any method that reconstructs a network based on its consistency with collections of such data will not be able to distinguish between networks N_1 and N_2. This includes all the methods whose data consists of clusters of taxa (e.g., [34]), triples (e.g., [35]), quartets (e.g., [36]), or any trees (e.g., [37]).

The same holds for many, sequence-based, maximum parsimony and maximum likelihood approaches proposed in recent papers. For maximum parsimony, a practical approach [2, 29–31] is to consider that the input is partitioned in a number of alignments $A_1, A_2, \ldots, A_m$, each from a different non-recombining genomic region (possibly consisting of just one site each), and then take, for each of these alignments, the best parsimony score $\mathbf{Ps}(T|A_i)$ among all those of the trees displayed by a network N. The parsimony score of N is then the sum of all the parsimony scores thus obtained. Formally, we have

$$\mathbf{Ps}(N|A_1, A_2, \ldots, A_m) = \sum_{i=1}^{m} \min_{T \in \mathcal{T}(N)} \mathbf{Ps}(T|A_i).$$

It is clear that if two networks display the same set of trees (as in Fig. 1), then their parsimony score with respect to any input alignments will be the same — because they take the minimum value over the same set $\mathcal{T}(N)$ — and thus they are indistinguishable to any method based on the maximum parsimony principle above.

As for maximum likelihood (ML), Nakhleh and collaborators [2, 32, 33, 38] have proposed an elegant framework whereby a phylogenetic network N is not only described by a network topology, but also edge lengths and inheritance probabilities associated to the reticulations of N. As a result, any tree T displayed by N has edge lengths — allowing the calculation of its likelihood $\mathbf{Pr}(A|T)$ with respect to any alignment A — and an associated probability of being observed $\mathbf{Pr}(T|N)$. The likelihood function with respect to a set of alignments $A_1, A_2, \ldots, A_m$, each from a different non-recombining genomic region, is then given by:

$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} \mathbf{Pr}(A_i|N) = \prod_{i=1}^{m} \sum_{T \in \mathcal{T}(N)} \mathbf{Pr}(A_i|T)\mathbf{Pr}(T|N).$$

Note that an important difference with the consistency-based and parsimony methods described above is that any tree T displayed by a network has now edge lengths and an associated probability $\mathbf{Pr}(T|N)$.

Unfortunately, this ML framework is also subject to identifiability problems. For example, it does not allow us to distinguish between networks with topologies N_1 and N_2 in Fig. 1: for every assignment of edge lengths and inheritance probabilities to N_1, there exist corresponding assignments to N_2 that make the resulting networks indistinguishable, that is, displaying the same trees, with the same edge lengths and the same probabilities of being observed (see the last section in the Supporting Information, S1 Text). As a result, the likelihoods of these two networks will be identical, regardless of the data, and no method

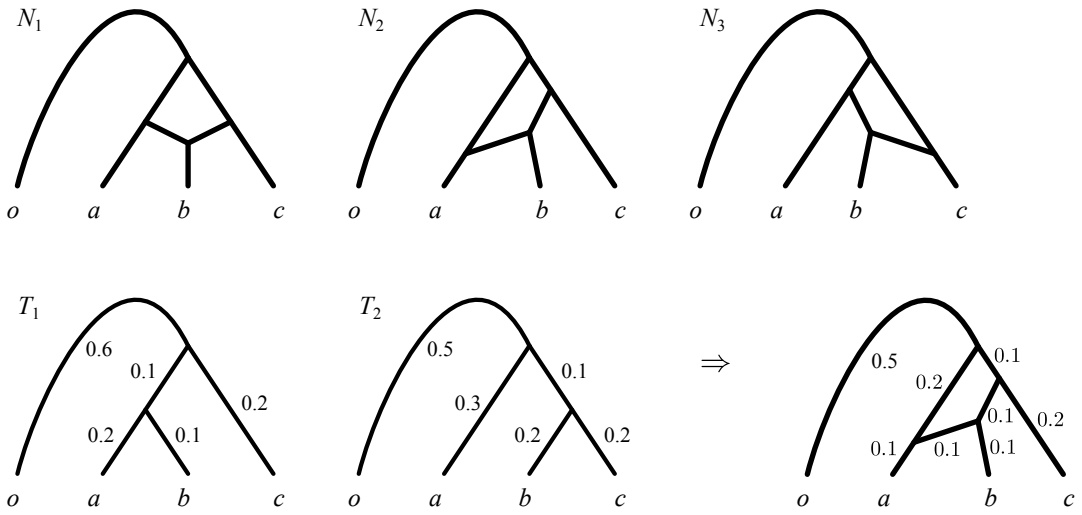


Fig 2. Edge lengths are informative to distinguish among different network topologies. The only network topology, among N_1, N_2 and N_3 that can display simultaneously T_1 and T_2 with the indicated edge lengths is N_2: see for example the edge lengths assignment in the bottom right corner.

based on this definition of likelihood will be able to favour one of them over the other. We refer to S1 Text for a more detailed discussion about networks with inheritance probabilities and likelihood-based reconstruction.

In general, we believe that these identifiability problems affect all network inference methods which seek consistency with unordered collections of sequence alignments or pre-inferred attributes such as clusters, triples, quartets or trees.

The Importance of Edge Lengths

In this paper, as in the ML framework above, we adopt networks and trees with edge lengths as the primary objects of our study. The primary motivation for this is that this choice makes our results directly relevant to the statistical approaches for network inference, all of which need edge lengths to measure the fit of a phylogeny with the available data. In addition to ML, these approaches include distance-based and Bayesian methods [39], which are also promising for future work.

However, there is another motivation for our choice: accounting for edge lengths solves some of the identifiability problems outlined above, as in some cases it allows to distinguish between networks with different topologies, which would be otherwise impossible to tell apart. For example, consider the three network topologies in Fig. 2 (top), where taxon o is an outgroup used to identify the root of the phylogeny for a, b and c. These networks show three very different evolutionary histories: in N_1 taxon b is the only one issued of a reticulation event — in other words the genome of b is recombinant — whereas in N_2 and N_3, it is a and c, respectively, that are recombinant. However, N_1, N_2 and N_3 display the same tree topologies — those of T_1 and T_2 — and thus would be indistinguishable to any approach that does not model edge lengths.

If instead edge lengths are accounted for (e.g. in a ML context) and the data supports T_1 and T_2 with the edge lengths in Fig. 2, then the only network fitting the data is N_2, with the edge lengths indicated at the bottom right. It is easy to check that N_2 now displays T_1 and T_2 with the shown edge lengths, whereas no edge length assignment to N_1 or N_3 can make these networks display T_1 and T_2.

We note that, throughout this paper, as in classical likelihood approaches, edge lengths measure

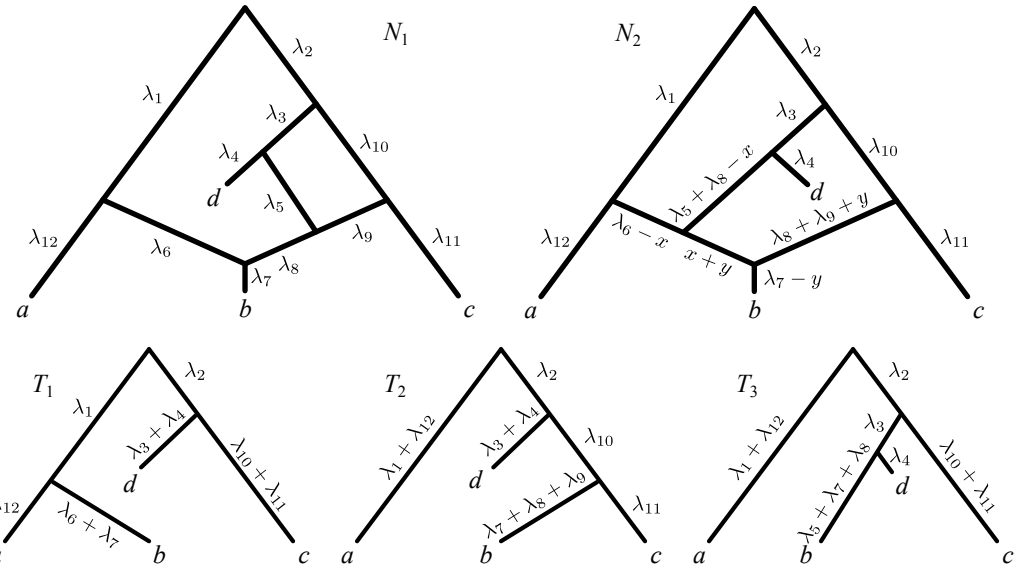


Fig 3. Indistinguishable networks. Two networks with edge lengths N_1, N_2 displaying the same set of trees $\mathcal{T}(N_1) = \mathcal{T}(N_2) = \{T_1, T_2, T_3\}$. For any choice of edge lengths $\lambda_1, \lambda_2, \ldots, \lambda_{12}$ for N_1, we define a family of edge length assignments for N_2, parameterized by x, y (with $-y < x < \min\{\lambda_6, \lambda_5 + \lambda_8\}$, $0 < y < \lambda_7$).

evolutionary divergence, for example in terms of expected number of substitutions per site. No molecular clock is assumed, meaning that we do not expect edge lengths to be proportional to time.

Remaining Identifiability Problems, and a Proposed Solution

Unfortunately, accounting for edge lengths only solves some of the identifiability problems for phylogenetic networks. Consider networks N_1 and N_2 in Fig. 3: for any set of edge lengths for N_1, there exist an infinity of edge length assignments for N_2 that make these two networks display exactly the same set of trees with the same edge lengths. In the following, we say that networks such as N_1 and N_2 are *indistinguishable*.

In fact it is not difficult to construct other examples of indistinguishable networks: each time a network has a reticulation v giving birth to only one edge (i.e. with outdegree 1), then we can reduce by $\Delta\lambda$ the length of this edge and correspondingly increase by $\Delta\lambda$ the lengths of the edges ending in v, without altering the set of trees displayed by the network. Note that this operation, which we refer to as "unzipping" reticulation v, can result in v coinciding with a speciation node or a leaf when $\Delta\lambda$ is taken to equal the length of the edge going out of v. For example in Fig. 3, one may fully unzip the two reticulation nodes in N_1, thus obtaining the network N' of Fig. 4. As expected, N_1 and N' display the same set of trees ($\{T_1, T_2, T_3\}$) and are thus indistinguishable. What is most interesting in this example is that, if we fully unzip the two reticulations in N_2 (the other network in Fig. 3, also displaying $\{T_1, T_2, T_3\}$), then we eventually end up obtaining N' again. As we shall see in the following, this is not a coincidence: the unzipping transformations described above lead to what we call the *canonical form* of a network; under mild assumptions, two networks are indistinguishable if and only if they have the same canonical form (e.g. N_1, N_2 in Fig. 3 have the same canonical form N'; formal definitions and statements in the Results section).

Here, we propose to deal with the identifiability issues for phylogenetic networks in the following

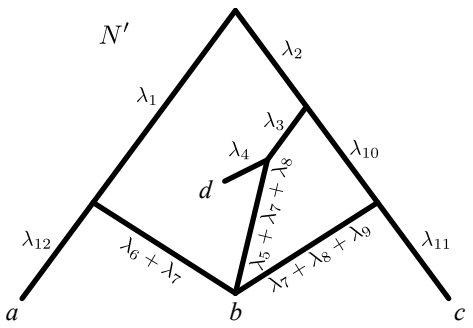


Fig 4. Canonical form of N_1 and N_2 in Fig. 3.

way: since no data will ever enable any of the standard inference methods described above to prefer a network over all of its indistinguishable equivalents, we propose that these methods *should only attempt to reconstruct what they can uniquely identify*, that is, networks in canonical form. This is a radical shift, not only for the developers of phylogenetic inference methods, who will see a drastic reduction of the solution space of their algorithms, but also for evolutionary biologists, who should abandon their hopes of seeing a network such as N_1 or N_2 in Fig. 3 being reconstructed by these inference methods.

Previous Work and Comparison

Limiting the scope of network reconstruction to topologically-constrained classes of networks has been a recurring theme and an important goal in the literature on phylogenetic networks. Examples of such classes include *galled trees* [40, 41], *galled networks* [42], *level-k networks* [43], *tree-child networks* [44], *tree-sibling networks* [45], networks with *visible reticulations* [1]. Although the ultimate goal should be to establish what can be inferred from biological data, most of the proposed definitions are computationally-motivated: in general the rationale behind these classes is the possibility of devising an efficient algorithm to solve some formalization of the reconstruction problem. None of these definitions claims to have biological significance.

Our goals are more basic: starting from the observation that not all phylogenetic networks are identifiable, since many of them are mutually indistinguishable with most inference approaches, we aim to define a class of networks that is (*existence* goal) large enough that every phylogenetic network has an equivalent (i.e. indistinguishable) network within this class and (*distinguishability* goal) small enough that no two networks within this class are indistinguishable. From our standpoint, the computationally-motivated definitions above are at the same time too broad and too restrictive. Too broad, because they determine a set of networks that includes many pairs of indistinguishable networks: for example the three indistinguishable networks in Fig. 2 are all galled trees — and thus belong to every single one of the classes mentioned above (which are all generalizations of galled trees). Too restrictive, because these classes of networks do not include simple networks that it should be possible to reconstruct from real data. For example, Fig. 5a shows a network N with edge lengths that is not tree-sibling, nor has the visible property, and thus is not galled, nor tree-child (for definitions, see [1]), but which in practice should be reconstructible: apart from the lengths of three edges (x, y, z), N is uniquely determined by the trees that it displays (a consequence of the formal results that we will show in the following), meaning that, given large amounts of data strongly supporting each of these (seven) trees with their correct edge lengths, any method for network inference properly accounting for edge lengths (e.g. based on ML) should be able to reconstruct N, or its canonical form N'.

To the best of our knowledge, only three classes of networks have claims of unique identifiability:

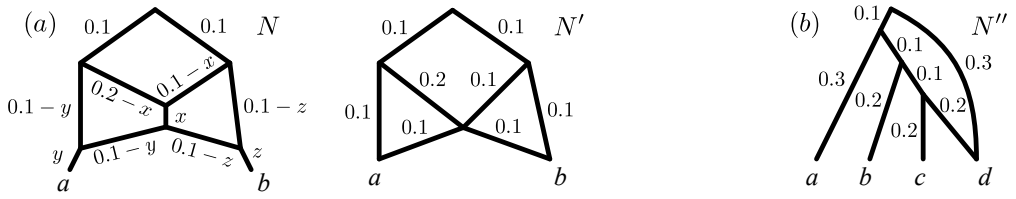


Fig 5. Examples of networks that can be uniquely recovered from the data they generate. (*a*) A network N, and its canonical form N', whose topologies are not galled trees, nor galled, tree-child, tree-sibling or regular networks, nor networks with visible reticulations. N, however, is uniquely determined by the trees it displays, with the exception of x, y and z, which can assume any value between 0 and 0.1. Because of the impossibility to determine these values, the canonical form N' has the corresponding edges collapsed. As N' is a network in canonical form satisfying the mild conditions of Corollary 2, N' is uniquely determined by the trees it displays. Note that N provides the biological interpretation for N'. (*b*) The network topology of N'' is such that there exists no regular network displaying the same set of (two) tree topologies as N''. Thus, restricting the scope of phylogenetic inference to regular networks would be very limiting. In our framework, N'' is a network in canonical form and thus uniquely determined by the trees it displays.

reduced networks [46,47], *regular* networks [48] and binary galled trees with no gall containing exactly 4 nodes [49]. These approaches bear some resemblances to ours, but do not include edge lengths in the definition of a network. Moreover, we argue that these classes of networks are still too narrow to be biologically relevant. We briefly describe and comment these previous works below.

Moret et al. [46] defined notions of reconstructible, indistinguishable and reduced networks that resemble concepts that we will introduce here. Although some of their results were flawed [47,50], some of the arguments in this introduction are inspired by their paper. Particularly relevant to the current paper is a reduction algorithm to transform a network into its *reduced version*. (However, the exact definition of the reduced version is unclear: as one of the authors later pointed out [47], "the reduction procedure of Moret et al. [46] is, in fact, inaccurate" and "in this paper we do not attempt to fix the procedure".) The concept of reduced version is analogous to that of canonical form here, as the authors claim that networks displaying the same tree topologies have the same reduced version (up to isomorphism; Theorem 2 in [46]). This is somehow a weaker analogue of one of our results (Corollary 1); weaker, because it does not claim that, conversely, networks with the same reduced version display the same tree topologies. To have an idea of the difference between our canonical form and the reduced version of Moret and colleagues, in Fig. 6 we compare the canonical form and the reduced version of the same network N_1. (N_1 and its reduced version are taken from Fig. 15 of [46] to avoid possible issues with the reduction algorithm.) As one can see, the canonical form retains more of the complexity of the original network.

Another reduction procedure on network topologies has been studied by Gambette and Huber [49], who prove that if two network topologies reduce to the same topology, then they must display the same tree topologies. Again, this is analogous to, but somehow weaker than our results, since it only provides a sufficient condition for networks to be indistinguishable (which in their context means to display the same tree topologies). This means that there can be irreducible networks that are indistinguishable (e.g. those in Fig. 2) thus failing to achieve the distinguishability goal. Moreover, Gambette and Huber [49] show that a particular class of network topologies (binary galled trees with no gall containing exactly 4 nodes) are uniquely identified by the tree topologies they display. It is clear that this class is too small to achieve the existence goal.

Finally, a regular network is a network topology N in which, among other requirements, no two distinct nodes have the same set of descendant leaves (see [48] for a formal definition and characterizations). This requirement implies, among other things, that N cannot contain any reticulation v with outdegree 1 (v and its direct descendant would have the same descendant leaves), which in turn implies that regular

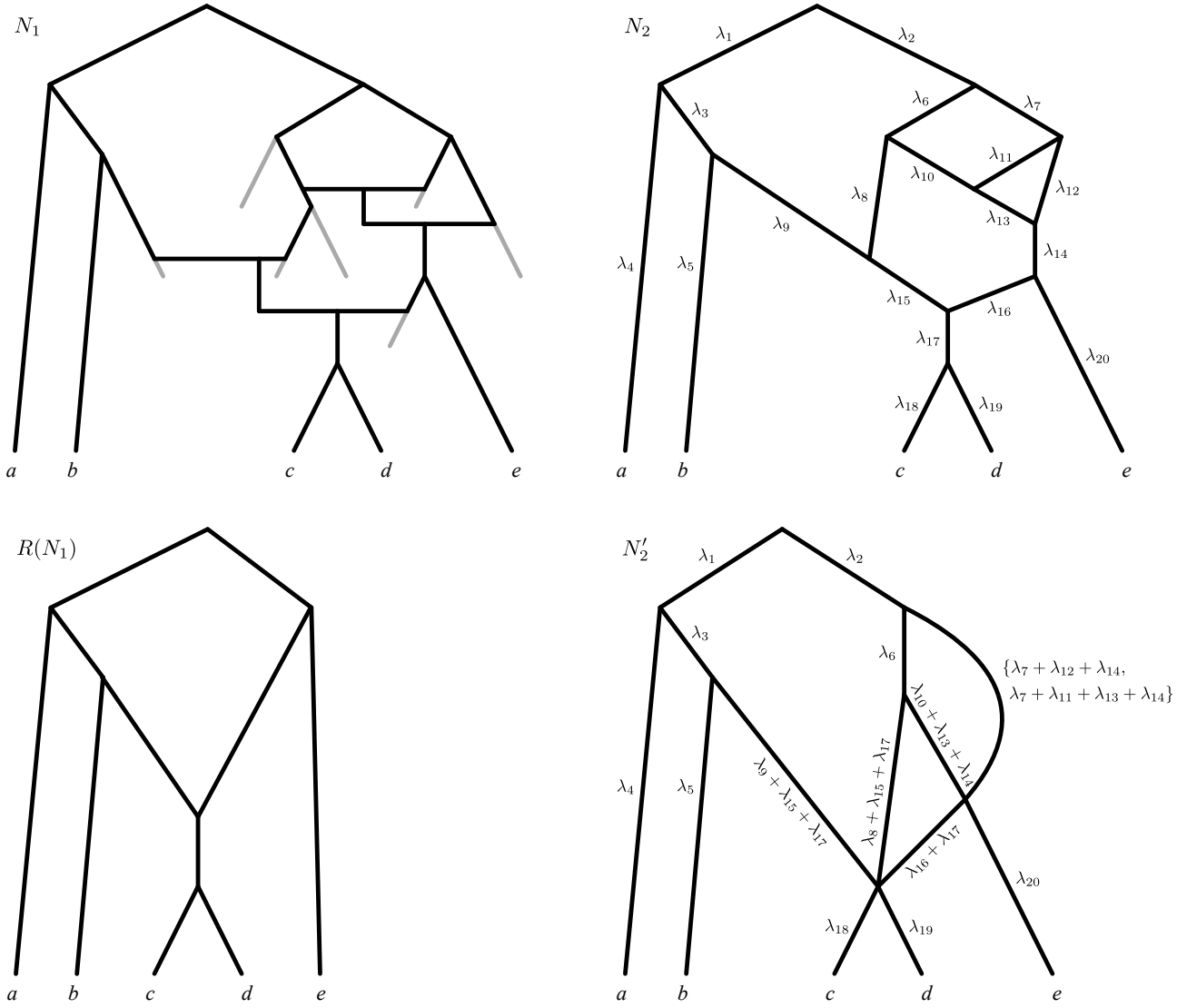


Fig 6. Comparison between the reduced version and the canonical form of a network. N_1 is the network topology in Fig. 15a of [46], where edges leading to extinct taxa are shown in grey, and reticulation events are represented by horizontal lines connecting the involved edges. N_2 is a phylogenetic network on the same set of taxa displaying the same evolutionary history, and showing edge lengths. $R(N_1)$ is the reduced version of N_1 (Fig. 15b of [46]). N_2' is the canonical form of N_2. Comparing $R(N_1)$ and N_2' reveals the difference in expressive power between reduced versions and canonical forms. Collapsing the edge above c and d in $R(N_1)$ yields the regular network displaying the same tree topologies as N_1 and N_2. Clearly, the reduced form $R(N_1)$ (and the regular form) retain less of the complexity of the original network N_1 than the canonical form N_2'. For example in $R(N_1)$ there remains no sign of the reticulate events ancestral to taxon e.

networks are special cases of our canonical networks (the latter however also specify edge lengths). In fact regular networks satisfy a property that is analogous to the one we prove here for canonical networks: a regular network N is uniquely determined by the tree topologies that it displays [51], meaning that there can be no other regular network N' displaying exactly the same set of tree topologies. Willson [51] shows this constructively by providing an algorithm that, given the (exponentially large) set of tree topologies displayed by a regular network R, reconstructs R itself. However, unlike for our canonical forms, for a given network there may exist no regular network displaying the same set of trees (e.g. consider the topology of N'' in Fig. 5b), thus failing to meet the existence goal. Regularity is in fact a very restrictive constraint for a network. For example, none of the networks in Fig. 5 and Fig. 7 is regular, despite the fact that their topologies are uniquely determined by the trees with edge lengths that they display (a consequence of our results further below). Finally, going back to Fig. 6, collapsing the edge above taxa c and d in $R(N_1)$ yields the regular network displaying the same tree topologies as N_1 and N_2. Again, this shows that the canonical form retains more of the complexity of the original network than its regular counterpart.

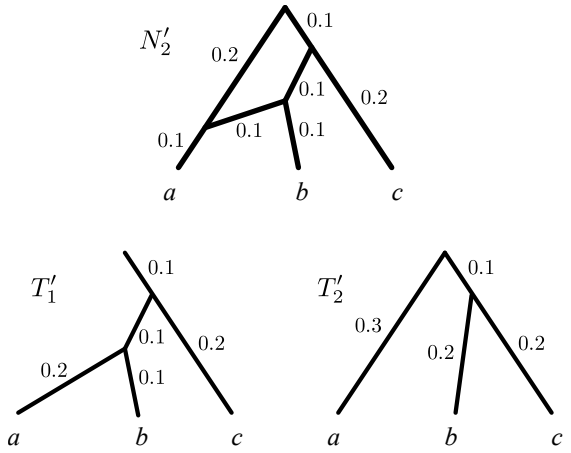


Fig 7. Trees displayed by a network. A rooted network N_2', and the trees it displays (T_1' and T_2'), obtained by removing a segment of length 0.5 from the outgroup lineage of N_2 in Fig. 2. In our formal setting, a network such as N_2 in Fig. 2 can either be represented as N_2' (by omitting the outgroup lineage, or part of it), or by rooting it in its outgroup (not shown).

Results

Our main result consists of formally proving that for every network N there exists a network N' in canonical form, indistinguishable from N; moreover, if we restrict ourselves to networks satisfying a mild condition (the NELP property below), such canonical form N' is unique (see Theorem 1). In other words, although in general a phylogenetic network N is not uniquely recoverable from the data it generates, there always exists a canonical version N' of N that is indeed determined by the data. Informally, N' is all that can be reconstructed about N.

In order to formally state this result, we here introduce a theoretical framework for explicit phylogenetic networks with branch lengths. A directed acyclic graph (DAG) is a simple directed graph that is free of directed cycles. A DAG is *rooted* if it contains precisely one node of indegree 0, called the *root*. All nodes of outdegree 0 in a DAG are called *leaves*. A *weighted rooted phylogenetic network* $N = (V, E, \varphi, \Lambda)$ on $\mathcal{X}$ (in this paper also called a *network* for simplicity) consists of a rooted DAG (V, E) whose leaves are bijectively labeled (via $\varphi : \mathcal{X} \to V$) with the elements of $\mathcal{X}$ (called *taxa*). Moreover, each edge $e \in E$ is associated to a set of positive weights, called *lengths*, $\Lambda(e) \subset \mathbb{R}_{>0}$. Figs. 3, 4, 5 contain examples of networks. A *reticulation* of a network N is a node $v \in V$ with indegree greater than 1. A *weighted phylogenetic tree* on $\mathcal{X}$ (a *tree* for simplicity) is a network on $\mathcal{X}$ with no reticulations and such that each edge e has a unique length ($|\Lambda(e)| = 1$), which we denote by $\lambda(e)$. Below, we discuss the biological justification of various aspects of the definitions above.

Let v be a node with indegree 1 and outdegree 1 in a tree. Node v is said to be *suppressible*. *Suppressing* v means removing the in-edge $e = (u, v)$ and the out-edge $f = (v, w)$ and then creating a new edge $g = (u, w)$ with length $\lambda(g) = \lambda(e) + \lambda(f)$. Let $N = (V, E, \varphi, \Lambda)$ be a network on $\mathcal{X}$. A *tree contained in* N is a tree $T = (V', E', \varphi', \lambda)$ on the same taxon set $\mathcal{X}$ such that: (1) the roots of T and N coincide, (2) the nodes and edges of T are also nodes and edges of N, that is $V' \subseteq V$ and $E' \subseteq E$, (3) taxon labels are unchanged, that is $\varphi' = \varphi$, and (4) the edge lengths of T are also edge lengths of N, that is, for every edge $e \in E'$, $\lambda(e) \in \Lambda(e)$. A *tree displayed by* N is a tree T' that can be obtained (up to isomorphism) by suppressing all suppressible nodes from a tree contained in N. The set of trees displayed by N is denoted by $\mathcal{T}(N)$. In Fig. 7, $\mathcal{T}(N_2')$ is the set of trees isomorphic to T_1' and T_2'. Two networks N_1 and N_2 are said to be *indistinguishable* if they display the same set of trees, that is $\mathcal{T}(N_1) = \mathcal{T}(N_2)$. For example, N_1 and N_2 in Fig. 3 are indistinguishable, as they display the same set of trees (T_1, T_2 and T_3, up to isomorphism).

Definition 1. Given a network N, a *funnel* is a node with indegree greater than 0 and outdegree 1. A *funnel-free* network, or *canonical* network, is a network that does not contain funnels. A *canonical form* of a network N is a network that is funnel-free and indistinguishable from N.

In Fig. 3, N_1 and N_2 each contain two funnels, and thus are not funnel-free. The network N' in Fig. 4 is a canonical form of N_1 and N_2 in Fig. 3, as N' is funnel-free and indistinguishable from N_1 and N_2. Similarly, N_2' in Fig. 6 is a canonical form of N_2. Note that nodes with indegree 1 and outdegree 1 are funnels. This implies that for trees the funnel-free condition coincides with the exclusion of suppressible nodes, which is a standard requirement in the definition of phylogenetic trees. It is thus appropriate to view the funnel-free condition as a natural extension of this requisite to networks.

Definition 2. A *weighted path* in a network $N = (V, E, \varphi, \Lambda)$ is a pair (π, λ), where π is a directed path in the graph (V, E) and λ is a function that associates each edge e in π with a length $\lambda(e) \in \Lambda(e)$. The *length* of a weighted path is the sum of the lengths assigned to its edges. A network satisfies the *NELP* (*no equally long paths*) property if no pair of distinct weighted paths having the same endpoints have the same length.

As we explain below, the NELP property is a mild condition to satisfy, unless edge lengths are taken to represent time. The following result states that if we restrict ourselves to networks satisfying the NELP

property, then every network has exactly one canonical form. An outline of its proof can be found in the Methods section, including an algorithm showing how to reduce a network to canonical form. The detailed proof is presented in S1 Text.

Theorem 1. *(i) Every network N has a canonical form. Moreover, (ii) if N has the NELP property, then there exists a unique canonical form of N among networks satisfying the NELP property (up to isomorphism).*

(The notion of isomorphism between networks is only used for mathematical rigor and is defined in S1 Text.) The following result provides a necessary and sufficient condition for two networks satisfying the NELP property to be indistinguishable.

Corollary 1. *Let N_1 and N_2 be networks with the NELP property and let N_1' and N_2' be their unique canonical forms satisfying the NELP property. Then N_1 and N_2 are indistinguishable if and only if N_1' and N_2' are the same network (up to isomorphism).*

The following result states that a canonical network with the NELP property is uniquely determined by the trees it displays:

Corollary 2. *Let N be a canonical network satisfying the NELP property. Then N is the unique (up to isomorphism) canonical network satisfying the NELP property that displays (all and only) the trees in $\mathcal{T}(N)$.*

We now discuss the biological significance of a number of technical aspects of our framework.

Definition of Networks and Trees Displayed by a Network

All the phylogenies considered here — trees or networks — are rooted. This is because we assume that the analysis uses an outgroup (possibly consisting of multiple taxa, and with no reticulations) for rooting. For simplicity, outgroup lineages are not included in our phylogenies (an exception to this is in Fig. 2). Note however that, because our phylogenies have edge lengths, and because omitting the outgroup is just a convention, the omitted lineages must have the same lengths for a network and all the trees it displays. For example, if we wish to omit the outgroup from N_2 in Fig. 2 and from the trees that it displays, then what we obtain are N_2', T_1' and T_2' in Fig. 7. This has a notable consequence: the trees displayed by a rooted network with edge lengths may have a root with outdegree 1 (e.g. T_1' in Fig. 7). For flexibility, we also allow a network to have a root with outdegree 1.

Moreover, we allow multiple lengths for an edge in a network, but not in a tree. For example, in Fig. 6, network N_2' has an edge with two lengths ($\lambda_7 + \lambda_{12} + \lambda_{14}$ and $\lambda_7 + \lambda_{11} + \lambda_{13} + \lambda_{14}$). The motivation behind multiple lengths lies in the observation that, whereas each edge in a phylogenetic tree describing the evolution of non-reticulating organisms trivially corresponds to a unique evolutionary path in the underlying real evolutionary history, when reticulate events have occurred this is not necessarily true: Fig. 8 and Fig. 9 show that some evolutionary scenarios can either be represented using multiedges (multiple edges with the same endpoints) or edges with multiple lengths. Although these two options are mathematically equivalent, graphically the second one leads to more compact representations, and this is why we choose to allow multiple lengths rather than multiedges. For our purposes we only need to consider the case where e has a finite set of lengths ($\Lambda(e) = \{\lambda_1(e), \ldots, \lambda_k(e)\}$).

Another unconventional aspect of our networks is the possibility of having nodes with in-degree and out-degree both greater than one. (See, e.g., the last common ancestor of c and d in N_2' in Fig. 6.) Traditionally, the internal nodes in a phylogenetic network are constrained to belong to one of two different categories: reticulate nodes, with more than one incoming edge and just one outgoing edge, and speciation (or coalescence) nodes, with one incoming edge and multiple outgoing edges. Because reticulate and speciation events are clearly distinct, it is reasonable to constrain internal nodes to only fall in the two

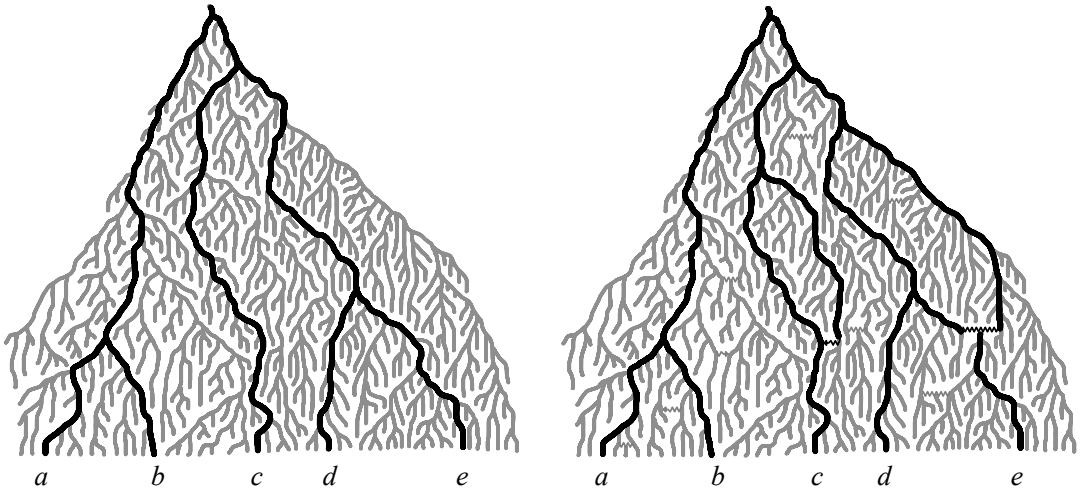


Fig 8. A non-reticulating evolutionary history (left) and a reticulating evolutionary history (right). The black lineages are those leading to a sampled set of taxa $\mathcal{X}$. The horizontal jagged lines represent reticulation events. Whereas representing the scenario on the left with a phylogenetic tree on $\mathcal{X}$ is straightforward, for the one on the right several options are possible. We show three alternative representations in Fig. 9.

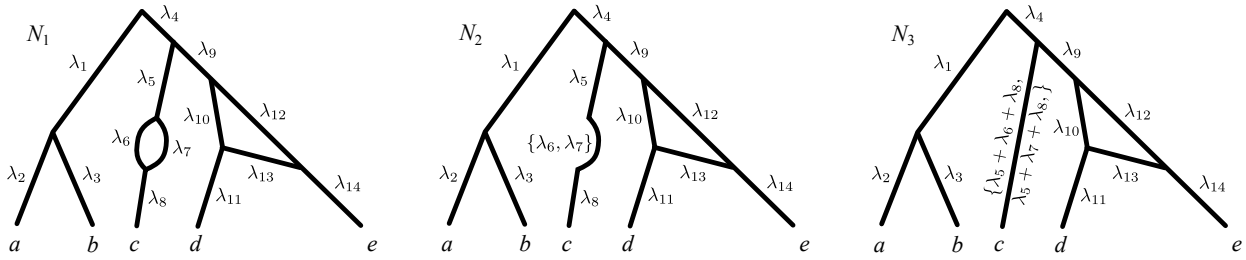


Fig 9. Alternative network representations for the evolutionary scenario in Fig. 8 (right). In our framework only N_2 and N_3 are networks.

categories above. In our framework, this requirement is dropped, and some networks, notably those in canonical form, may have nodes that both represent reticulate and speciation events. In this case, it is important to understand that these nodes represent a potentially complex (and unrecoverable) reticulate scenario, followed by one or more speciation events. Compare, for example, network N and its canonical form N' in Fig. 5, or N_2 and N_2' in Fig. 6. (In the latter, it is especially instructive to consider the reticulate history above the direct ancestor of taxon e.)

The NELP Property

We use network N_1 of Fig. 3 to illustrate the NELP property. In N_1 there are three distinct weighted paths having as endpoints the root of N_1 and the direct ancestor of b. The lengths of these paths are

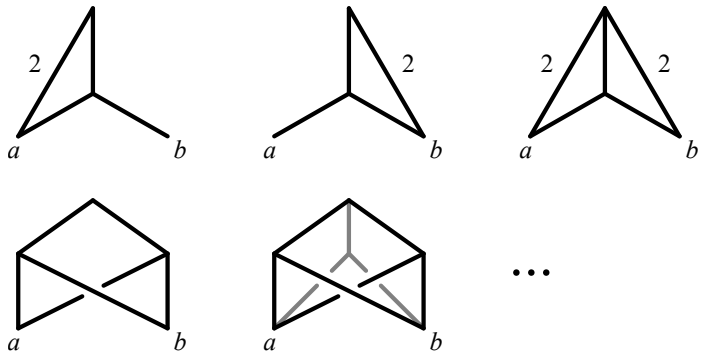


Fig 10. Different (non-isomorphic) but indistinguishable funnel-free networks. All edges are assumed to have the (unique) length 1 unless otherwise displayed. These networks do not satisfy the NELP property, showing that this is a necessary condition for the uniqueness of canonical forms (Theorem 1(ii)). The ellipsis at the end represents the fact that an infinite number of such networks can be obtained by adding any number of copies of the subgraph in grey in the last network.

$\ell_1 = \lambda_1 + \lambda_6$, $\ell_2 = \lambda_2 + \lambda_3 + \lambda_5 + \lambda_8$ and $\ell_3 = \lambda_2 + \lambda_{10} + \lambda_9 + \lambda_8$. Moreover, there is another pair of paths having the same endpoints: those of lengths $\ell_4 = \lambda_3 + \lambda_5$ and $\ell_5 = \lambda_{10} + \lambda_9$. Thus N_1 has the NELP property if and only if the three numbers ℓ_1, ℓ_2 and ℓ_3 are all different (note that this implies that also ℓ_4 and ℓ_5 are different). If edge lengths are taken to represent evolutionary change, rather than time, this is a very mild requirement: when edge lengths are drawn at random from a continuous distribution, the probability that two paths get exactly the same length is zero.

On the other hand, the NELP property does not hold for phylogenetic networks where edge lengths are taken to represent time. For these networks, canonical forms may not be unique (see Fig. 10 for an example of this). Even in this case, we believe that inference methods should only consider phylogenetic networks in their canonical form, as this allows to reduce the solution space without any loss in "expressive power": since every network N has (at least one) canonical form that displays exactly the same set of trees — and therefore has the same fit with the data as N — restricting the solution space to canonical forms always leaves at least one optimal network within this space. The real weakness of using canonical forms in a molecular clock context is that if a canonical form is not unique, then it cannot be considered representative of all the networks indistinguishable from it. As an example of this, consider the indistinguishable networks in Fig. 10: none of these is representative of all the others.

Discussion

Our results are both negative and positive. The bad news is that any method that scores the fit between a network N and the available data — which may be sequences, distances, splits, trees (with or without edge lengths) — based on the set of trees displayed by N must face an important theoretical limitation: regardless of the amount of available data from the taxa under consideration, some parts of the network representing their evolutionary history may be impossible to recover — most notably the relative order of consecutive reticulate events (see, e.g., Fig. 3). The good news is that, when edge lengths are taken into account, we can set precise limits to what is recoverable: the canonical form of a network N is

a simplified version of N that excludes all the unrecoverable aspects of N. In a canonical form, reticulate events are brought as forward in time as possible, causing the collapse of multiple consecutive nodes. (Compare network N_2 and its canonical form N_2' in Fig. 6.) The importance of the canonical form N' of a network N lies in the fact that, if we restrict our consideration to networks with the NELP property, N' is the unique canonical network consistent with perfect and unlimited data from the taxa in N.

There is an interesting analogy between soft polytomies in classical phylogenetics and collapsed nodes in a canonical network. Both represent lack of knowledge about the order of evolutionary events: speciations or more generally lineage splits in the first case, and reticulate events in the second. However, there is also an important difference between them: whereas in principle polytomies can be resolved by collecting further data from the taxa in the tree (for example, by extensive sequencing of their genomes [52]), the standard network inference methods considered here cannot resolve collapsed nodes in a canonical network, *irrespective of the amount of data from the taxa under consideration.* This difference is mitigated by the observation that increased taxon sampling may indeed permit to resolve the collapsed nodes, when the new lineages break adjacencies between reticulate nodes. However, such lineages may not always exist or they may be difficult to sample.

The present work has several consequences that should be of interest both to the biologists concerned by the use of methods for phylogenetic network inference, and to the researchers interested in the development of these methods. We illustrate these consequences starting from a well-known problem of network inference methods, that of multiple optima. It has been noted before that many of the inference methods that have been recently proposed — especially those solely based on topological features — often return multiple optimal networks: Huson and Scornavacca show a striking example of this (Fig. 2 in [53]), where the problem of finding the simplest network displaying two given tree topologies admits at least 486 optimal solutions.

The existence of multiple optimal networks for a given data set is essentially due to two reasons: *insufficient data* and *non-identifiability*. For the example of 486 optimal solutions, this large number may be partly due to the fact that the goal was to achieve consistency with only two tree topologies. More data may enable to discriminate among the 486 returned networks. Non-identifiability, which occurs when none of the allowed data can discriminate between two or more networks, is a more serious problem than insufficient data, as it cannot be solved by simply increasing the size of the input sample. Another interesting example appears in a paper by Albrecht et al. [54], which we reproduce here in Fig. 11. Here, there are only three optimal networks, essentially differing for which of the three clades $\{A.bicornis,$ $A.longissima,$ $A.sharonensis\}$, $\{A.uniaristata,$ $A.comosa\}$ and $\{A.tauschii\}$ is considered as a hybrid (in this example reticulations represent hybridizations). This pattern is entirely analogous to that of the three networks in Fig. 2 (with a, b and c replaced by the three clades above), meaning that these three networks are indistinguishable to methods not accounting for edge lengths. Therefore, in this example, the existence of multiple optimal solutions is *entirely* due to non-identifiability.

All this motivates three recommendations:

1. It is important to use data in a way that causes non-identifiability to be as limited as possible. For example, as we have seen, accounting for edge lengths solves some cases of non-identifiability (e.g., in Fig. 2) although it does not eliminate this problem altogether (e.g., in Fig. 3).

2. Given an inferred network $\hat{N}$, it is important to know the set of networks that are theoretically impossible to distinguish from $\hat{N}$: no matter the amount of data, they will all receive the same support as $\hat{N}$. We may call this set the *indistinguishable class* of $\hat{N}$. The biologist using an inference method must be aware that $\hat{N}$ is not the only network supported by the data.

3. It would be highly useful to devise inference methods that instead of searching for (or directly constructing) solutions in the space of all possible networks, only considers one element per indistinguishable class. This has the potential to significantly speed up the inference.

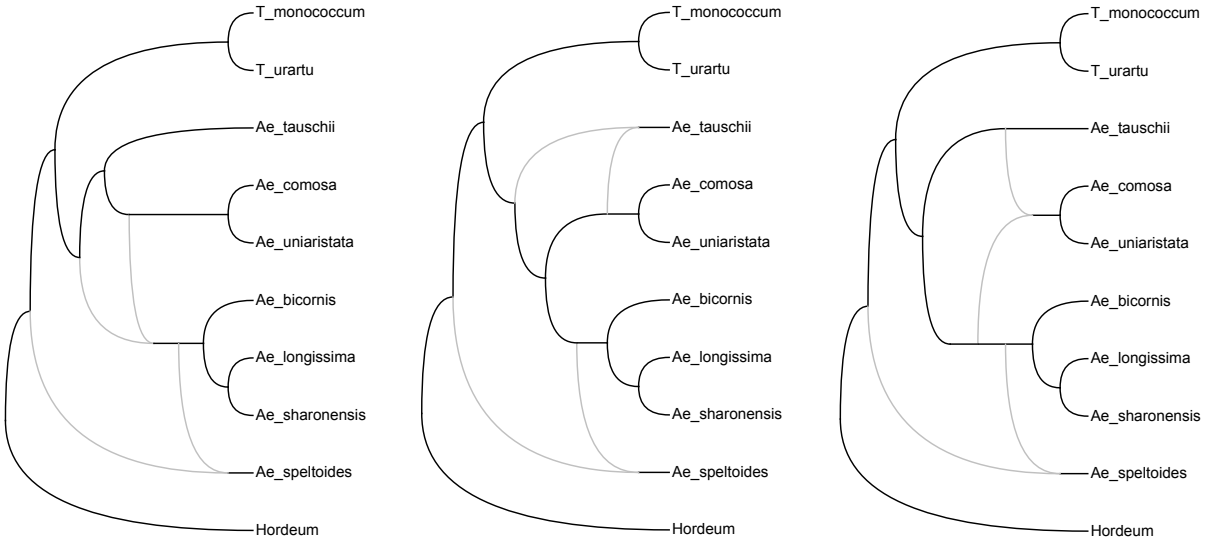


Fig 11. Real-world example of indistinguishable network topologies. (Reproduced from [54], Fig. 4.) Three network topologies that display the two tree topologies in Fig. 3 of [54]. Note that these three networks are analogous to N_1, N_2 and N_3 in Fig. 2 of the current paper: they each contain a reticulation cycle with three outgoing edges leading to the same three clades: {*A.bicornis, A.longissima, A.sharonensis*}, {*A.uniaristata, A.comosa*} and {*A.tauschii*} (in Fig. 2 instead of three clades we have three taxa a, b and c).

Correspondingly, we recommend that edge lengths should be accounted for in the analyses (point 1) and, for each of the indistinguishable classes resulting from this choice, we identify a canonical network that, for all practical purposes, can be considered to be unique. Most important to the end users, we propose that a canonical network $\hat{N}$ is what should be given as the result of the inference, with the caveat that $\hat{N}$ is a way to represent a class of networks that are all equally supported (point 2). In a canonical form $\hat{N}$, the aspects that are not common to all networks in this class are collapsed, as described above. This will help the evolutionary biologist to locate the uncertainties in the phylogeny, and possibly to choose further taxa to resolve them. Finally, we propose that inference methods only attempt to search among — or construct — phylogenetic networks in their canonical form (point 3).

We note that accounting for yet more characteristics of the data may reduce (or eliminate altogether) the identifiability issues for phylogenetic networks. In the case of sequence-based methods, one may take into account the natural order of sites within a sequence [11–13, 55, 56]. Similarly, for reconstruction methods based on collections of subtrees, one could observe and use the relative position of the different genomic regions supporting the input trees. However, these relative positions must be conserved across the genomes being analyzed, a condition which may hold for recombining organisms (e.g. individuals within a population or different viral strains), but which is not obvious when studying a group of taxa that have undergone reticulate events (e.g., hybridization) at some point in a distant past.

The main conclusion of the present study is the following: unless one abandons any optimization criterion that scores a network solely based on the trees it displays, the reconstruction should be carried out in a reduced space of networks: that of the canonical forms defined here. The motivation for this lies in the fact that canonical networks are guaranteed to be uniquely determined, if sufficient data are available. Once a canonical form $\hat{N}$ is inferred, it must be kept in mind that even assuming that the inference is free of statistical error, the true phylogenetic network is just one of the many networks having $\hat{N}$ as canonical form. Compared to what biologists are used to for phylogenetic trees — where in principle it is always possible to resolve uncertainties — it is clear that this requires an important change of perspective.

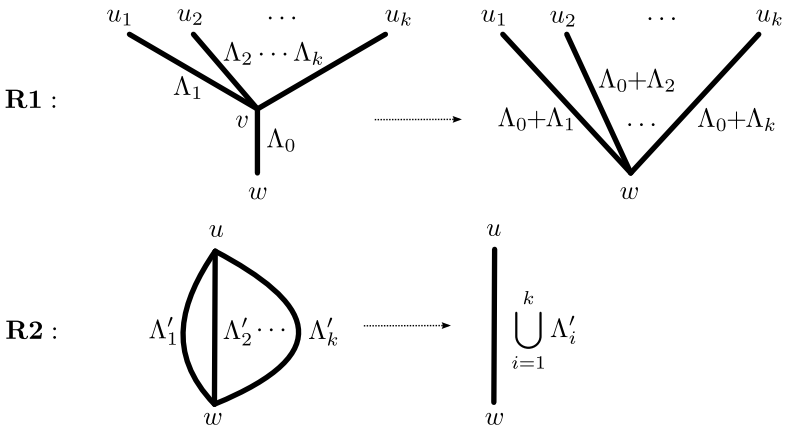


Fig 12. The two rules at the basis of the canonical reduction algorithm.

Methods

The following three subsections describe the proofs of Theorem 1 part (i), of Theorem 1 part (ii), and of their corollaries, respectively. In the case of Theorem 1 part (ii), only the gist of the proof is provided here. The proof in full detail is deferred to S1 Text.

Reduction Algorithm

In order to prove that any network N has a canonical form, we describe an algorithm to transform N into a canonical network indistinguishable from N. The algorithm simply consists of repeatedly applying to $N = (V, E, \varphi, \Lambda)$ one of the following two reduction rules, until neither can be executed (see Fig. 12):

Funnel suppression (R1) given a funnel v with $k \geq 1$ in-edges $(u_1, v), (u_2, v), \ldots, (u_k, v)$ and out-edge (v, w), remove v and all these edges from N and introduce k new edges $(u_1, w), (u_2, w), \ldots, (u_k, w)$. For all $i \in \{1, 2, \ldots, k\}$ assign to (u_i, w) the lengths $\Lambda((u_i, w)) := \Lambda((u_i, v)) + \Lambda((v, w))$, where the sum of two sets of numbers A and B is defined as $A + B = \{a + b : a \in A, b \in B\}$.

Multiedge merging (R2) given a collection of multi-edges (u, w) with multiplicity k and lengths $\Lambda'_1, \Lambda'_2, \ldots, \Lambda'_k$, replace these edges with a single edge with lengths $\bigcup_{i=1}^{k} \Lambda'_i$.

An example of the reduction of a network to its canonical form is shown in Fig. 13. Note that, even if the algorithm may temporarily produce multi-edges, the network produced in the end obviously does not have any multi-edge (otherwise we could still apply rule R2).

Proof of part (i) of Theorem 1. We must prove that any network $N = (V, E, \varphi, \Lambda)$ has a canonical form. For this, we apply the reduction algorithm described above, thus obtaining a sequence $N_0 = N, N_1, \ldots, N_m$, where each N_{i+1} is obtained from N_i by applying either R1 or R2. Neither R1 nor R2 can be applied to N_m. We prove that N_m is a canonical form of N. Although, strictly speaking, N_i may not be a network (as it potentially contains multi-edges), the notion of trees displayed by N_i, and thus that of indistinguishability, trivially extends to these multigraphs.

First, note that the algorithm terminates after a finite number of iterations (m). This is true because at each iteration the size of E is reduced by at least one. Moreover, the resulting network N_m is funnel-free, since no reduction of type R1 can be applied to it.

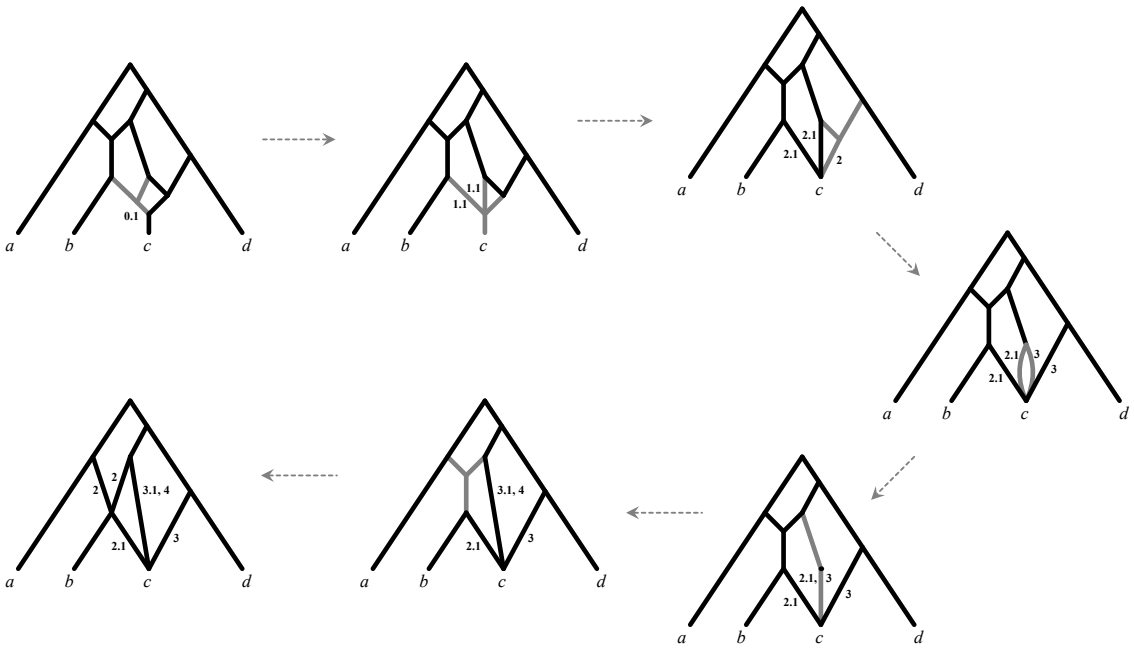


Fig 13. Reduction of a network to its canonical form. Gray edges are those to which the next reduction rule is applied. All edges are assumed to have the (unique) length 1 unless otherwise displayed.

What is left to prove is that N_m is indistinguishable from $N = N_0$. To this end we prove that, at each iteration, N_i and N_{i+1} are indistinguishable, i.e. $\mathcal{T}(N_i) = \mathcal{T}(N_{i+1})$. In other words any tree T is displayed by N_i if and only if T is displayed by N_{i+1}.

Let T be displayed by N_i. Then T can be obtained by suppressing all suppressible nodes from a tree T_i contained in N_i. We consider three cases. (1) If none of the edges in T_i is involved in the reduction transforming N_i into N_{i+1}, then clearly T_i is still contained in N_{i+1} and thus T is still displayed by N_{i+1}. (2) If T_i is involved in a R1 reduction, then it contains a funnel v and it contains one of the in-edges of the funnel, say (u_j, v), with length $\lambda_j \in \Lambda_j = \Lambda((u_j, v))$, along with the out-edge (v, w), with length $\lambda_0 \in \Lambda_0 = \Lambda((v, w))$. Now, let T_{i+1} be the tree obtained from T_i by suppressing the suppressible node v and thus creating a new edge (u_j, w) with length $\lambda_j + \lambda_0$. Because the R1 reduction creates a new edge (u_j, w) with length set $\Lambda_j + \Lambda_0$, containing the value $\lambda_j + \lambda_0$, then T_{i+1} is contained in N_{i+1}. Moreover, it easy to see that T can still be obtained by suppressing all suppressible nodes from T_{i+1}. Thus T is still displayed by N_{i+1}. (3) If T_i is involved in a R2 reduction, then it contains one of the edges of a multi-edge (u, w), with a length λ belonging to one of the length sets $\Lambda'_1, \Lambda'_2, \ldots, \Lambda'_k$ associated to the k copies of (u, w). Thus we have that $\lambda \in \bigcup_{i=1}^{k} \Lambda'_i$, which implies that T_i is still contained in N_{i+1} and thus T is still displayed by N_{i+1}. This concludes the proof of $\mathcal{T}(N_i) \subseteq \mathcal{T}(N_{i+1})$.

In order to prove that, conversely, $\mathcal{T}(N_{i+1}) \subseteq \mathcal{T}(N_i)$, one can proceed in a similar way as above: if T is displayed by N_{i+1}, then T can be obtained by suppressing all suppressible nodes from a tree T_{i+1} contained in N_i. By considering three cases analogous to the ones above regarding the involvement of T_{i+1} in the reduction transforming N_i into N_{i+1}, we can prove that in all these cases T is already displayed by N_i. Thus N_i and N_{i+1} are indistinguishable, which concludes our proof. □

We note informally that the order of application of the possible reductions in the algorithm above is irrelevant to the end result. To see this, it suffices to show that if two different reductions are applicable

to a network, then the result of applying them is the same irrespective of the order of application. As we do not need this remark for the other results in this paper, we do not give a formal proof of it.

Lemma 1. *Let N be a network and N' a canonical form of N obtained by applying the reduction algorithm. If N satisfies the NELP property, then N' satisfies the NELP property.*

Proof. We prove that for each basic step of the reduction algorithm — transforming N_i into N_{i+1} via a reduction rule R1/R2 — if N_i satisfies the NELP property, then N_{i+1} also satisfies it. Suppose the contrary; then, N_{i+1} contains two distinct weighted paths ρ_1, ρ_2 with the same endpoints u and v and same lengths. Because R1/R2 cannot create new nodes, u and v are also nodes in N_i. Moreover, it is easy to see that each weighted path ρ in N_i from u to v gives rise to exactly one weighted path $f(\rho)$ in N_{i+1} from u to v, with exactly the same length as ρ. Now take two weighted paths in N_i, one in the preimage $f^{-1}(\rho_1)$ and the other in the preimage $f^{-1}(\rho_2)$. These two weighted paths in N_i are distinct (as $\rho_1 \neq \rho_2$), have the same endpoints (u and v) and the same length. But then N_i violates the NELP property, leading to a contradiction. We thus have that if N_i satisfies the NELP property, then N_{i+1} also satisfies it. By iterating the argument above for each step in the reduction algorithm, the lemma follows. □

Uniqueness of the Canonical Form for Networks Satisfying the NELP

The proof of Theorem 1, part (ii), is rather technical. In this section, we introduce a number of new concepts and state the main intermediate results that are necessary to obtain this result. We leave their detailed proofs to S1 Text, together with the obvious definitions of basic concepts such as that of *isomorphic networks*, *sub-network* and *union* of two networks.

Definition 3. (*Root-leaf path, prefix, postfix, wishbone, crack.*) Let N be a network on $\mathcal{X}$ and (π, λ) be a weighted path in N from the root of N to a leaf labelled by $x \in \mathcal{X}$. Now consider the sub-network $P = (V(\pi), E(\pi), \varphi|_{\{x\}}, \lambda)$ on $\{x\}$ consisting of all the nodes and edges in π and associated labels. Any sub-network of N such as P is called a *root-leaf path* of N. Given a root-leaf path P and a node v belonging to it, any weighted path formed by all the ancestors [descendants] of v in P is a *prefix* [*suffix*] of P. Note that a prefix [suffix] only consists of one node when v is the root [leaf] of P. A *wishbone* of N is any sub-network of N formed by taking the union of two root-leaf paths that have in common only a prefix. A *crack* of N is any sub-network of N formed by taking the union of two root-leaf paths that have in common only a prefix and a suffix.

Fig. 14 illustrates the definitions above. Note that any root-leaf path P is both a wishbone and a crack, as P is the result of the union of P with itself, and P has a common prefix and a common suffix with P. Moreover, any sub-network R that can be obtained from a root-leaf path by attributing two lengths to one of its edges e is a crack. Finally, note that wishbones and cracks are networks, and thus the notion of isomorphism (Definition 5 in S1 Text) can be applied to them.

The proof of part (ii) in Theorem 1 depends on two important results (Propositions 1 and 2 below), whose proofs can be found in S1 Text. The first states that a network with the NELP property is uniquely determined by the wishbones and cracks it contains.

Proposition 1. *Two networks N_1 and N_2 with the NELP property are isomorphic if and only if they contain the same wishbones and cracks (up to isomorphism).*

Proposition 1 is interesting on its own as it suggests an enumerative algorithm to verify whether two networks with the NELP property are isomorphic. Unfortunately this algorithm would be impractical, as the number of wishbones (or cracks) in a network is not polynomial in the size of the network. Also note that we require N_1 and N_2 to satisfy the NELP property because there exist non-isomorphic networks containing the same wishbones and cracks: for example the networks in the bottom line of Fig. 10. The second result that we need is the following:

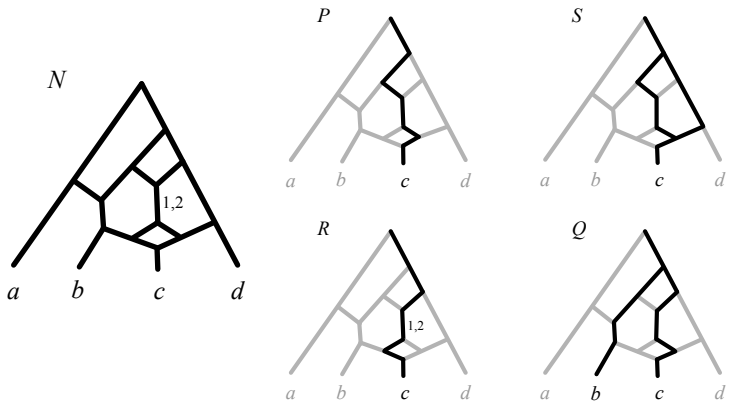


Fig 14. Illustration of Definition 3. P (edges in black) is a root-leaf path of N and thus both a wishbone and a crack of N. R and S (black) are cracks of N. Q (black) is a wishbone of N. All edges are assumed to have the (unique) length 1 unless otherwise displayed.

Proposition 2. *Let N_1 and N_2 be two indistinguishable funnel-free networks, satisfying the NELP property. Then they contain the same wishbones and cracks (up to isomorphism).*

Proof of part (ii) of Theorem 1. Let N be a network with the NELP property and N' a canonical form of N obtained by applying the reduction algorithm. By Lemma 1, N' satisfies the NELP property. Now suppose that there exists another canonical form of N, called N'', satisfying the NELP property. By transitivity, N' and N'' are indistinguishable. Because N' and N'' are indistinguishable, funnel-free and with the NELP property, N' and N'' must contain the same wishbones and cracks (because of Proposition 2). But then, because of Proposition 1, N' and N'' are isomorphic. □

We note that some of our arguments in S1 Text lead us to conjecture that a funnel-free network satisfying the NELP property cannot be indistinguishable from a funnel-free network violating the NELP property. This claim would allow us to simplify the statement of Theorem 1: networks with the NELP property would be guaranteed to have a unique canonical form (not just among networks with the NELP property, but among *all* networks). Unfortunately, to this date, we were unable to prove this conjecture. Nonetheless, note that the reduction algorithm returns, for any network with the NELP property, its *unique* canonical form with the NELP property (by Lemma 1).

Corollaries

It remains to prove the two corollaries at the end of the Results section. The first one states that two networks N_1 and N_2 satisfying the NELP property are indistinguishable if and only if their unique canonical forms with the NELP property, N_1' and N_2' respectively, are isomorphic. By Lemma 1, N_1' and N_2' can be obtained by applying the reduction algorithm to N_1 and N_2.

Proof of Corollary 1. The *if* part trivially follows from the transitivity of indistinguishability. As for the *only if* part, note that (again by transitivity) N_1' is indistinguishable from N_2. As it is also funnel-free, N_1' is a canonical form of N_2. Because N_2 can only have one canonical form satisfying the NELP property (by Theorem 1(ii)), N_1' and N_2' must be the same network (up to isomorphism). □

As for Corollary 2, we recall that it states that a canonical network N with the NELP property is uniquely determined by the trees it displays.

Proof of Corollary 2. Let N and N' be indistinguishable canonical networks satisfying the NELP property. Then, N and N' are both canonical forms of N satisfying the NELP. But then, by Theorem 1(ii), N and N' must be the same network (up to isomorphism). □

Acknowledgments

We are grateful to O.Gascuel for advice on the structure of the paper.

References

1. Huson DH, Rupp R, Scornavacca C (2011) Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press.

2. Nakhleh L (2011) Evolutionary phylogenetic networks: models and issues. In: The Problem Solving Handbook in Computational Biology and Bioinformatics, Springer. pp. 125–158.

3. Morrison DA (2011) Introduction to Phylogenetic Networks. RJR Productions.

4. Mallet J (2007) Hybrid speciation. Nature 446: 279–283.

5. Nolte AW, Tautz D (2010) Understanding the onset of hybrid speciation. Trends in Genetics 26: 54–58.

6. Ochman H, Lawrence J, Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

7. Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society B: Biological Sciences 277: 819-827.

8. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122–1125.

9. Rambaut A, Posada D, Crandall K, Holmes E (2004) The causes and consequences of HIV evolution. Nature Reviews Genetics 5: 52–61.

10. Simon-Loriere E, Holmes EC (2011) Why do RNA viruses recombine? Nature Reviews Microbiology 9: 617–626.

11. Song YS, Hein J (2005) Constructing minimal ancestral recombination graphs. Journal of Computational Biology 12: 147-169.

12. Minichiello M, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. American Journal of Human Genetics 79: 910–922.

13. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genetics 10: e1004342.

14. Huson D, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23: 254–267.

15. Bryant D, Moulton V (2004) Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution 21: 255-265.

16. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.

17. Hallström BM, Kullberg M, Nilsson MA, Janke A (2007) Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. Molecular Biology and Evolution 24: 2059–2068.

18. Lorentz Center (2012). The future of phylogenetic networks. Available: `http://www.lorentzcenter.nl/lc/web/2012/515/description.php3?wsid=515`. Accessed 20 Oct 2014.

19. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, et al. (2013) Networks: expanding evolutionary thinking. Trends in Genetics 29: 439–441.

20. Morrison D (2013). What are evolutionary networks currently used for? Available: `http://phylonetworks.blogspot.fr/2013/10/what-are-evolutionary-networks.html`. Accessed 20 Oct 2014.

21. Delwiche CF, Palmer JD (1996) Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Molecular Biology and Evolution 13: 873–882.

22. Morgan DR (2003) nrDNA external transcribed spacer (ETS) sequence data, reticulate evolution, and the systematics of Machaeranthera (Asteraceae). Systematic Botany 28: 179–190.

23. Marhold K, Lihová J (2006) Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Systematics and Evolution 259: 143–174.

24. Koblmüller S, Duftner N, Sefc KM, Aibara M, Stipacek M, et al. (2007) Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika – the result of repeated introgressive hybridization. BMC Evolutionary Biology 7: 7.

25. Richards TA, Soanes DM, Foster PG, Leonard G, Thornton CR, et al. (2009) Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. The Plant Cell 21: 1897–1911.

26. Dyer RJ, Savolainen V, Schneider H (2012) Apomixis and reticulate evolution in the Asplenium monanthes fern complex. Annals of Botany 110: 1515–1529.

27. Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biology and Evolution 4: 466–485.

28. Huson D, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. Genome Biology and Evolution 3: 23.

29. Jin G, Nakhleh L, Snir S, Tuller T (2006) Efficient parsimony-based methods for phylogenetic network reconstruction. In: Proceedings of the 5th European Conference on Computational Biology (ECCB). volume 23 of *Bioinformatics*, pp. e123-e128.

30. Jin G, Nakhleh L, Snir S, Tuller T (2007) Inferring phylogenetic networks by the maximum parsimony criterion: A case study. Molecular Biology and Evolution 24: 324-337.

31. Jin G, Nakhleh L, Snir S, Tuller T (2009) Parsimony score of phylogenetic networks: hardness results and a linear-time heuristic. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6: 495–505.

32. Jin G, Nakhleh L, Snir S, Tuller T (2006) Maximum likelihood of phylogenetic networks. Bioinformatics 22: 2604-2611.

33. Park HJ, Nakhleh L (2012) Inference of reticulate evolutionary histories by maximum likelihood: the performance of information criteria. BMC Bioinformatics 13: S12.

34. van Iersel L, Kelk S, Rupp R, Huson D (2010) Phylogenetic networks do not need to be complex: Using fewer reticulations to represent conflicting clusters. In: Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB). volume 26 of *Bioinformatics*, pp. i124-i131.

35. To TH, Habib M (2009) Level-k phylogenetic networks are constructable from a dense triplet set in polynomial time. In: Combinatorial Pattern Matching: Proceeding of the 20th Annual Symposium Combinatorial Pattern Matching (CPM). volume 5577 of *LNCS*, pp. 275-288.

36. Grünewald S, Forslund K, Dress A, Moulton V (2007) Qnet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. Molecular Biology and Evolution 24: 532–538.

37. Baroni M, Semple C, Steel M (2006) Hybrids in real time. Systematic Biology 55: 46–56.

38. Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS genetics 8: e1002660.

39. Radice R (2011) A Bayesian Approach to Phylogenetic Networks. Ph.D. thesis, University of Bath.

40. Gusfield D, Eddhu S, Langley C (2003) Efficient reconstruction of phylogenetic networks with constrained recombinations. In: Proceedings of the IEEE Computer Society Conference on Bioinformatics (CSB). IEEE Computer Society, p. 363.

41. Wang L, Zhang K, Zhang L (2001) Perfect phylogenetic networks with recombination. Journal of Computational Biology 8: 69–78.

42. Huson DH, Rupp R, Berry V, Gambette P, Paul C (2009) Computing galled networks from real data. Bioinformatics 25: i85-i93.

43. Choy C, Jansson J, Sadakane K, Sung WK (2005) Computing the maximum agreement of phylogenetic networks. Theoretical Computer Science 335: 93-107.

44. Cardona G, Rosselló F, Valiente G (2007) Comparison of tree-child phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6: 552-569.

45. Cardona G, Llabrés M, Rosselló F, Valiente G (2008) A distance metric for a class of tree-sibling phylogenetic networks. Bioinformatics 24: 1481-1488.

46. Moret B, Nakhleh L, Warnow T, Linder C, Tholse A, et al. (2004) Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE Transactions on Computational Biology and Bioinformatics 1: 13–23.

47. Nakhleh L (2010) A metric on the space of reduced phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7: 218–222.

48. Baroni M, Semple C, Steel MA (2004) A framework for representing reticulate evolution. Annals of Combinatorics 8: 391-408.

49. Gambette P, Huber KT (2012) On encodings of phylogenetic networks of bounded level. Journal of Mathematical Biology 65: 157–180.

50. Cardona G, Rosselló F, Valiente G (2008) Tripartitions do not always discriminate phylogenetic networks. Mathematical Biosciences 211: 356–370.

51. Willson SJ (2011) Regular networks can be uniquely constructed from their trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8: 785–796.

52. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361–375.

53. Huson DH, Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61: 1061-1067.

54. Albrecht B, Scornavacca C, Cenci A, Huson DH (2012) Fast computation of minimum hybridization networks. Bioinformatics 28: 191-197.

55. Hein J (1993) A heuristic method to reconstruct the history of sequences subject to recombination. Journal of Molecular Evolution 36: 396-405.

56. Snir S, Tuller T (2009) The Net-HMM approach: Phylogenetic network inference by combining maximum likelihood and hidden Markov models. Journal of Bioinformatics and Computational Biology 7: 625–644.

Supporting Information Legends

S1 Text. Supporting Information: a mathematical theory of explicit phylogenetic networks with edge lengths. This document provides an introduction to the mathematical theory of explicit phylogenetic networks with edge lengths, leading in particular to the proofs of Propositions 1 and 2, which are necessary for the proof of Theorem 1, part (ii). In the last section, we consider networks with inheritance probabilities and their relevance for likelihood-based reconstruction.

A mathematical theory of explicit phylogenetic networks with edge lengths.

This document is structured in six sections, in which we develop a theory of phylogenetic networks with edge lengths. The first section introduces the notion of isomorphism between such networks and states some obvious propositions; the second section looks in more detail at the process whereby a network displays a tree; the third shows characterizations for both the NELP and the funnel-free property; the fourth and fifth sections derive the proofs of the two propositions that are necessary for the proof of Theorem 1, part (ii). The last section shows an example relevant to likelihood frameworks modelling inheritance probabilities, in addition to edge lengths. The notation and definitions introduced in the Results section and in the Methods section within the paper will be used here. We recall in particular that, throughout this paper, networks are rooted DAGs whose leaves are bijectively labeled by taxa and whose edges have a finite set of strictly positive lengths.

Isomorphisms and sub-networks

Definition 4. Given a network $N = (V, E, \varphi, \Lambda)$, a *sub-network* $N' = (V', E', \varphi', \Lambda')$ of N is any 4-tuple such that: (a) N' is a network, (b) $V' \subseteq V$, (c) $E' \subseteq V'^2 \cap E$, (d) φ' is the restriction of φ to the taxa associated to the leaves of (V', E') and, (e) for every edge $e \in E'$, $\Lambda'(e) \subseteq \Lambda(e)$. The *union* $N_1 \cup N_2$ of two sub-networks of N, $N_1 = (V_1, E_1, \varphi_1, \Lambda_1)$, $N_2 = (V_2, E_2, \varphi_2, \Lambda_2)$, having the same root, is the sub-network $N' = (V_1 \cup V_2, E_1 \cup E_2, \varphi', \Lambda')$, where φ' is the restriction of φ to the taxa associated to the leaves of $(V_1 \cup V_2, E_1 \cup E_2)$ and for every edge $e \in E_1 \cup E_2$, $\Lambda'(e) = \Lambda_1(e) \cup \Lambda_2(e)$ (where we take the liberty to let $\Lambda_i(e) = \emptyset$ whenever $e \notin E_i$).

Note that we define the union for sub-networks of N sharing the same root, because this ensures that such union is still a network. Although these requirements could be relaxed, the definition above is sufficient for the purposes of the current paper.

Definition 5. Let $N_1 = (V_1, E_1, \varphi_1, \Lambda_1)$ and $N_2 = (V_2, E_2, \varphi_2, \Lambda_2)$ be two networks on $\mathcal{X}$. N_1 and N_2 are *isomorphic* if there exists a bijection $f : V_1 \to V_2$ (called an *isomorphism*) such that:

(i) for every $x \in \mathcal{X}$, $f(\varphi_1(x)) = \varphi_2(x)$;

(ii) for every pair of nodes $(u, v) \in V_1^2$, $(u, v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2$ and whenever both these edges exist, $\Lambda_1((u, v)) = \Lambda_2((f(u), f(v)))$.

The following four lemmas are trivially true, and for brevity we do not include their proofs here.

Lemma 2. *Let N_1 and N_2 be isomorphic networks. Then every sub-network of N_1 is isomorphic to some sub-network of N_2.*

Definition 6. Let P be a root-leaf path of a network N. The *depth* of a node v in P is the length of the weighted path in P from the root of P to v. We say that P is *to* x, if P is a network on the set $\{x\}$, that is, if its only leaf is labelled by taxon x.

Lemma 3. *A root-leaf path P_1 is isomorphic to P_2 if and only if (a) P_2 is a root-leaf path to the same taxon as P_1, and (b) for every node v_i in any of the two root-leaf paths, say P_i, there exist a node v_j in the other root-leaf path, P_j, such that the depth of v_i in P_i is equal to the depth of v_j in P_j.*

Lemma 4. *A wishbone $W = P_1 \cup P_2$ is isomorphic to W' if and only if W' is a wishbone and can be written as $W' = P_1' \cup P_2'$, so that (a) the longest common prefix of P_1 and P_2 has the same length as the longest common prefix of P_1' and P_2', and (b) P_1' and P_2' are isomorphic to P_1 and P_2, respectively.*

Lemma 5. *A crack $K = P_1 \cup P_2$ is isomorphic to K' if and only if K' is a crack and can be written as $K' = P_1' \cup P_2'$, so that (a) the longest common prefix of P_1 and P_2 has the same length as the longest common prefix of P_1' and P_2', (b) the longest common suffix of P_1 and P_2 has the same length as the longest common suffix of P_1' and P_2', and (c) P_1' and P_2' are isomorphic to P_1 and P_2, respectively.*

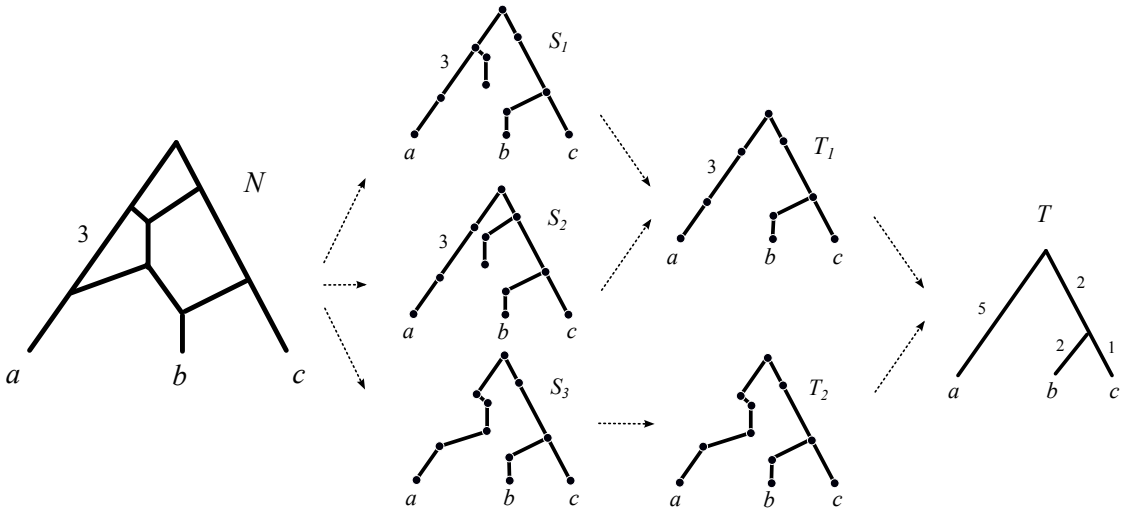


Figure S1. Illustration of the definitions of *switching* of a network, *tree contained* in a network, *tree displayed* by a network and *embeddings*. N is a network, S_1, S_2 and S_3 are three of its switchings, T_1 and T_2 are two trees contained in N, T is a tree displayed by N, and T_1 and T_2 are two embeddings of T in N. Nodes are explicitly shown in S_1, S_2, S_3, T_1 and T_2 for clarity. Unless otherwise shown, all edges are assumed to have length 1.

Switchings, trees weakly displayed and embeddings

Definition 7. Let $N = (V, E, \varphi, \Lambda)$ be a network. A *switching* $S = (V, E', \varphi, \lambda)$ of N is obtained from N by doing the following:
(i) for each node $v \in V$, delete all incoming edges of v except one; let E' be the resulting set of edges;
(ii) for each edge $e \in E'$, assign to e a single length $\lambda(e) \in \Lambda(e)$.

Note that technically a switching is not a network, as action (i) above may create leaves in S that are not labelled by any taxon. Biologically, a switching corresponds to the evolutionary tree describing the history of a single (indivisible and thus non-recombining) character carried by the root of N; some lineages in this tree may never reach any leaf of N. Moreover, since each edge in this tree corresponds to a unique path in the underlying real evolutionary history, it is clear why a switching only allows one length per edge (see Fig. 8 in the main text). Note that the trees contained by a network (defined in the Results section within the paper) can also be seen as the trees that can be obtained from a switching by removing all nodes and edges that have no descendant labelled by a taxon.

Definition 8. Let N be a network. A tree *weakly contained* in N is a tree T that is a sub-network of N and has the same root as N. A tree *weakly displayed* by N is any tree T can be obtained (up to isomorphism) by suppressing all suppressible nodes from a tree T' weakly contained in N. Tree T' is called an *embedding* of T in N. The set of trees weakly displayed by N is denoted by $\tilde{\mathcal{T}}(N)$.

The difference with the definition of trees displayed by a network is that a network on $\mathcal{X}$ can only display trees on $\mathcal{X}$, whereas it can weakly display any tree on $\mathcal{X}'$, with $\emptyset \subset \mathcal{X}' \subseteq \mathcal{X}$. Note that if a tree T is displayed by a network N, then T is also weakly displayed by N. The definitions of switching of a network, tree contained in a network, tree displayed by a network and embeddings are illustrated in Fig. S1. We note that every switching of a network N gives rise to a unique tree contained in N and

every tree contained in N gives rise to a unique tree displayed by N. However, the converses of these two propositions are not true: several switchings of N can give rise the same tree contained in N, and several trees contained in N can give rise to the same tree displayed by N. The latter means that $T \in \mathcal{T}(N)$ can have several embeddings in N.

Lemma 6. *Let T be a tree on $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ with no suppressible nodes. Let $P_1, P_2, \ldots, P_n$ be the root-leaf paths in T to $x_1, x_2, \ldots, x_n$, respectively. For all $i, j \in \{1, 2, \ldots, n\}$, let λ_i be the length of P_i and let λ_{ij} be the length of the longest common prefix of P_i and P_j. Any embedding of T is the union of n root-leaf paths $P'_1, P'_2, \ldots, P'_n$ to $x_1, x_2, \ldots, x_n$, such that for all $i, j \in \{1, 2, \ldots, n\}$, P'_i has length λ_i, $P'_i \cup P'_j$ is a wishbone and the length of the longest common prefix of P'_i and P'_j is λ_{ij}.*

Proof. Let T_e be an embedding of T. By definition of embedding, suppressing all suppressible nodes in T_e gives rise to a tree T' isomorphic to T. Because suppressing nodes does not change the taxon set of a tree, T_e is a tree on the same taxa as T', and thus on the same taxa as T, that is $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. Because T_e is a tree on $\mathcal{X}$, T_e equals the union of its n root-leaf paths to $x_1, x_2, \ldots, x_n$ — which we call $P'_1, P'_2, \ldots, P'_n$ respectively — and each pair of these paths have nothing in common other than a prefix, meaning that $P'_i \cup P'_j$ is a wishbone. Now let λ'_i denote the length of P'_i and λ'_{ij} denote the length of the longest common prefix of P'_i and P'_j. It remains to prove that $\lambda'_i = \lambda_i$ and $\lambda'_{ij} = \lambda_{ij}$, for all $i, j \in \{1, 2, \ldots, n\}$.

Because suppressing nodes does not change the length of a root-leaf path nor the taxon labelling its leaf, then the root-leaf path to x_i in T', which is derived from P'_i, must have length λ'_i. But because T and T' are isomorphic, then they contain the same root-leaf paths, up to isomorphism (by Lemma 2 and Lemma 3). The (unique) root-leaf paths to x_i in T and T' are thus isomorphic and have therefore the same length, that is, $\lambda_i = \lambda'_i$.

Similarly, because suppressing nodes transforms a wishbone in another wishbone, without changing the length of the longest prefix common to its two root-leaf paths, nor the taxa labelling their leaves, then the root-leaf paths to x_i and x_j in T', which are derived from P'_i and P'_j, respectively, must form a wishbone W', with a longest common prefix of length λ'_{ij}. Because T and T' are isomorphic, then they contain the same wishbones, up to isomorphism (by Lemma 2 and Lemma 4), meaning that W' and $P_i \cup P_j$ must be isomorphic. By Lemma 4, the longest common prefix of P_i and P_j must then have length $\lambda_{ij} = \lambda'_{ij}$. □

Lemma 7. *Let T, T' denote trees and N, N' denote networks.*

(a) If N and N' are isomorphic, then $\tilde{\mathcal{T}}(N) = \tilde{\mathcal{T}}(N')$.
(b) If T' is obtained by suppressing all suppressible nodes from T, then $\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T')$.
(c) If T' is an embedding of T, then $\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T')$.
(d) If T is weakly contained in N, then $\tilde{\mathcal{T}}(T) \subseteq \tilde{\mathcal{T}}(N)$.
(e) If T is weakly displayed by N, then $\tilde{\mathcal{T}}(T) \subseteq \tilde{\mathcal{T}}(N)$.

Proof. **(a)** Let T be a tree weakly contained in N. Let f be an isomorphism between N and N'. It is easy to see that the restriction of f to the nodes of T defines an isomorphism between T and a tree T' that is weakly contained in N', meaning that N and N' must weakly contain the same trees (up to isomorphism). Therefore N and N' must weakly display the same trees. **(b)** Let T, and thus T', be trees on $\mathcal{Y}$. Given a nonempty subset $\mathcal{X} \subseteq \mathcal{Y}$, let $T_{\mathcal{X}}$ be the (unique) tree on $\mathcal{X}$ weakly contained in T, and let $T'_{\mathcal{X}}$ be the (unique) tree on $\mathcal{X}$ weakly contained in T'. It is easy to see that $T'_{\mathcal{X}}$ can be obtained by suppressing some suppressible nodes from $T_{\mathcal{X}}$. It follows that the tree obtained by suppressing all suppressible nodes from $T'_{\mathcal{X}}$ is the same as that obtained by suppressing all suppressible nodes from $T_{\mathcal{X}}$. This means that a tree on $\mathcal{X}$ is weakly displayed by T if and only if it is weakly displayed by T'. Since this is true for any nonempty $\mathcal{X} \subseteq \mathcal{Y}$, point (b) follows. **(c)** By definition, if T' is an embedding of T, then there exist a tree T_i, isomorphic to T, that can be obtained by suppressing all suppressible nodes from T'. But then, by point (a), $\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T_i)$ and, by point (b), $\tilde{\mathcal{T}}(T_i) = \tilde{\mathcal{T}}(T')$. By transitivity, $\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T')$. **(d)** If T is weakly contained in N, then every tree weakly contained in T is also weakly contained in N. Therefore

every tree weakly displayed by T is also weakly displayed by N. **(e)** If T is weakly displayed by N, then by definition there exist an embedding T' of T in N. But then, by point (c), $\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T')$, and, because T' is weakly contained in N, $\tilde{\mathcal{T}}(T') \subseteq \tilde{\mathcal{T}}(N)$ (by point(d)). By transitivity, $\tilde{\mathcal{T}}(T) \subseteq \tilde{\mathcal{T}}(N)$. □

Lemma 8. *For every tree T_{wc} weakly contained in a network N, there exists a tree T_c contained in N that weakly contains T_{wc}.*

Proof. Let N be a network on $\mathcal{X}$ and T_{wc} be a tree on $\mathcal{X}' \subseteq \mathcal{X}$. Number the taxa in $\mathcal{X} \setminus \mathcal{X}'$ so that we have $\mathcal{X} \setminus \mathcal{X}' = \{x_1, x_2, \ldots, x_{|\mathcal{X}|-|\mathcal{X}'|}\}$. For convenience of notation, let $T_0 = T_{wc}$. For every $i \in \{1, 2, \ldots, |\mathcal{X}| - |\mathcal{X}'|\}$, we now show how to define T_i on $\mathcal{X}' \cup \{x_1, \ldots, x_i\}$ that weakly contains T_{i-1} and is weakly contained in N. Let P_i be a root-leaf path in N composed by the edges traversed by the following walk from the leaf labelled by x_i to the root in N: from the leaf labelled by x_i always take an edge e of N in its inverse direction (from its head to its tail), with any of e's associated lengths, until you end up in a node v belonging to T_{i-1}; from then on, follow the (inverse) path from v to the root of T_{i-1} (which coincides with that of N). By construction, P_i and T_{i-1} share a common prefix and nothing else. If we now define T_i as the union of T_{i-1} and P_i, it is clear that T_i is a tree on $\mathcal{X}' \cup \{x_1, \ldots, x_i\}$ that weakly contains T_{i-1} and is weakly contained in N. Now consider the sequence of trees $T_{wc} = T_0, T_1, \ldots, T_{|\mathcal{X}|-|\mathcal{X}'|}$. Each of these trees is weakly contained in N, and (because the relation of weak containment is transitive) weakly containing all its predecessors. Because $T_{|\mathcal{X}|-|\mathcal{X}'|}$ is a tree on $\mathcal{X}$ weakly contained in N, $T_{|\mathcal{X}|-|\mathcal{X}'|}$ is contained in N and weakly contains T_{wc}, thus concluding the proof. □

Proposition 3. *Let N and N' be networks. Then they are indistinguishable if and only if they weakly display the same trees.*

Proof. The *if* part is trivial. Suppose N and N' weakly display the same trees. Then they must be networks on the same taxon set $\mathcal{X}$, otherwise the network on the larger set of taxa, say, N, would weakly display trees with taxa that can never be present in trees weakly displayed by N'. Because N and N' weakly display the same trees on all taxon subsets $\mathcal{X}' \subseteq \mathcal{X}$, then they also display the same trees on $\mathcal{X}$, that is, they are indistinguishable.

As for the *only if* part, we prove that, assuming N and N' are indistinguishable, $\tilde{\mathcal{T}}(N) = \tilde{\mathcal{T}}(N')$. We just prove $\tilde{\mathcal{T}}(N) \subseteq \tilde{\mathcal{T}}(N')$, as the proof of $\tilde{\mathcal{T}}(N') \subseteq \tilde{\mathcal{T}}(N)$ is symmetric. Let $T \in \tilde{\mathcal{T}}(N)$ be a tree weakly displayed by N, and let T_e be an embedding of T in N. Because T_e is weakly contained in N, by Lemma 8 there exists a tree T_c contained in N that weakly contains T_e. But then, by points (c) and (d) in Lemma 7,

$$\tilde{\mathcal{T}}(T) = \tilde{\mathcal{T}}(T_e) \subseteq \tilde{\mathcal{T}}(T_c). \tag{1}$$

Now let T_d be the tree obtained by suppressing all suppressible nodes in T_c. Because T_c is contained in N, then T_d is displayed by N and thus by N', given that N and N' are indistinguishable. Therefore, by points (b) and (e) in Lemma 7,

$$\tilde{\mathcal{T}}(T_c) = \tilde{\mathcal{T}}(T_d) \subseteq \tilde{\mathcal{T}}(N'). \tag{2}$$

By putting together relations (1) and (2), we have that $T \in \tilde{\mathcal{T}}(T) \subseteq \tilde{\mathcal{T}}(N')$. Because this holds for any $T \in \tilde{\mathcal{T}}(N)$, we conclude $\tilde{\mathcal{T}}(N) \subseteq \tilde{\mathcal{T}}(N')$. □

Corollary 3. *Let N and N' be indistinguishable networks. Then, for every root-leaf path P in N, there exists in N' an equally long root-leaf path to the same taxon as P.*

Proof. Suppose there exists in N a root-leaf path to x of length λ. By suppressing all suppressible nodes from this path, one obtains a tree T consisting of one edge of length λ, whose head is labelled by x. Because N and N' are indistinguishable, they both weakly display T (by Proposition 3). But since any embedding of T must consist of a single root-leaf path to x of length λ (by Lemma 6), such a root-leaf path must exist in N'. □

Characterizing the funnel-free and the NELP properties

In this section we investigate what it means for a network to be canonical and to satisfy the NELP property. We start by characterizing the topological constraint (absence of funnels) that defines canonical networks.

Proposition 4. *A network N is funnel-free if and only if, for any two distinct non-root nodes u and w in N, there exist two node-disjoint directed paths from u and w to two distinct leaves in N.*

Proof. First, note that if N is not funnel-free, then the stated property does not hold: it suffices to take a funnel as u and its only direct descendant as w and then it is clear that any two paths from u and w cannot be node-disjoint.

Second, we prove that in a funnel-free network, for any two distinct nodes u and w, neither of which coincides with the root, there always exist two node-disjoint paths π_{ux} and π_{wy}, from u to x and from w to y respectively, where x and y are leaves of N. Let $d_N(v)$ denote the number of proper descendants (that is, not including v) of a node v in a network N. We prove our claim by induction on $d_N(u) + d_N(w)$.

First, suppose $d_N(u) + d_N(w) = 0$. In this case u and w are leaves and the thesis trivially holds.

Then, suppose $d_N(u) + d_N(w) = n > 0$. Assume, without loss of generality, that $d_N(w) \geq d_N(u)$. Then $d_N(w) > 0$, which means that w is an internal node. As a consequence, w must have at least two children nodes, because otherwise w would be a funnel or a root with outdegree 1 (which we have excluded). Of these children nodes, at least one must be different from u. Call this node w'. Because w' is a descendant of w, then $d_N(w') < d_N(w)$ and so $d_N(u) + d_N(w') < n$. By the inductive hypothesis, we can then assume that the thesis holds for the pair of (different, non-root) nodes (u, w'). That is, there exist in N two disjoint directed paths π_{ux} and $\pi_{w'y}$, from u to x and from w' to y respectively, such that x and y are leaves of N. Now form a new path π_{wy} by appending the edge (w, w') at the beginning of $\pi_{w'y}$. Note that w cannot be part of π_{ux} because otherwise w would be a descendant of u and that would contradict $d_N(w) \geq d_N(u)$. Because of this, and because $\pi_{w'y}$ and π_{ux} are disjoint, then also π_{wy} and π_{ux} are disjoint. The thesis then holds also in the inductive step and the theorem follows. □

We now examine the NELP property. It turns out that this property is equivalent to requiring the uniqueness of embeddings for all trees weakly displayed by the network. In order to prove this, we show the following trivial result that will be useful throughout this document.

Lemma 9. *A network N satisfies the NELP property if and only if, for every taxon x in N, all root-leaf paths in N to x have different lengths.*

Proof. For the *if* part, note that if N does not satisfy the NELP property, then there are two distinct weighted paths (π_1, λ_1), (π_2, λ_2) with the same lengths and endpoints. In this case it is easy to extend these paths to two distinct root-leaf paths P_1 and P_2 leading to the same taxon x and having the same length. As for the *only if* part, the existence of two root-leaf paths in N of equal lengths to the same taxon x clearly implies that N violates the NELP property. □

Proposition 5. *A network N satisfies the NELP property if and only if every tree weakly displayed by N has a unique embedding in N.*

Proof. For the *only if* part, let T be a tree on $\{x_1, x_2, \ldots, x_n\}$, weakly displayed by N, with $\lambda_1, \lambda_2, \ldots, \lambda_n$ being the lengths of the root-leaf paths to $x_1, x_2, \ldots, x_n$ in T. Suppose T_1 and T_2 are two embeddings of T in N. Then T_1 and T_2 must each consists of the union of n root-leaf paths to $x_1, x_2, \ldots, x_n$ of lengths $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively (Lemma 6). But because N satisfies the NELP property, then for any $i \in \{1, 2, \ldots, n\}$ there can be only one path to x_i of length λ_i (Lemma 9), meaning that $T_1 = T_2$.

As for the *if* part, suppose every tree displayed by a network N has a unique embedding in N, but that N does not satisfy the NELP property. Then (by Lemma 9) there exist two distinct root-leaf paths P_1 and P_2 to the same taxon (say, x) that have the same length λ. But then P_1 and P_2 are distinct

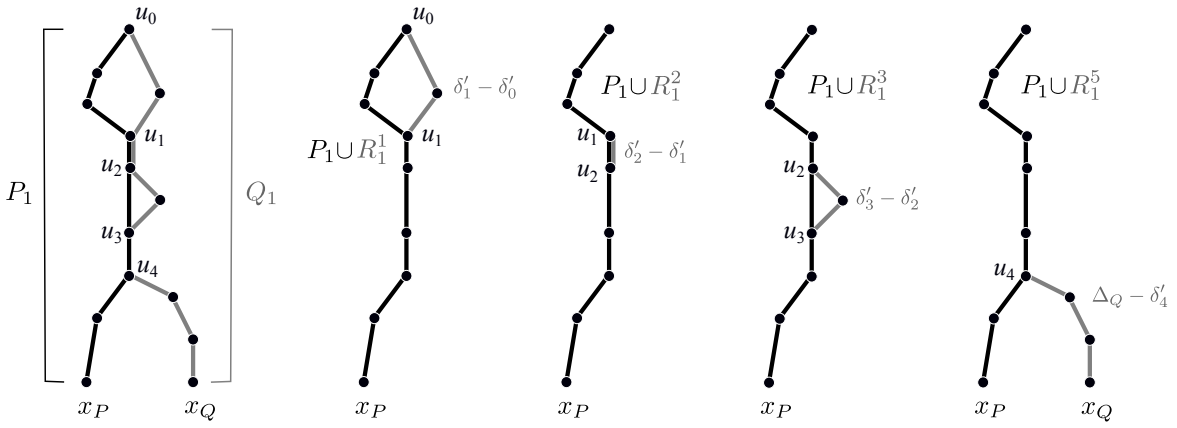


Figure S2. Two root-leaf paths P_1 and Q_1 intersecting in 5 nodes, their corresponding cracks $P_1 \cup R_1^1, P_1 \cup R_1^2, P_1 \cup R_1^3$ and wishbone $P_1 \cup R_1^5$ (see the proof of Lemma 10).

embeddings in N of the tree consisting of a single edge whose head is labelled by x, thus contradicting the uniqueness of embeddings. □

Note that the network in Fig. S1, which has several embeddings for T (namely T_1 and T_2), violates the NELP property: for example there are three different paths of length 5 from the root to the leaf labelled by a. Also note that the uniqueness of embeddings for the trees in $\mathcal{T}(N)$ does not imply the NELP property: consider for example the top-left network in Fig. 10 in the main text, where the two trees displayed by this network have unique embeddings, but the network does not satisfy the NELP property.

Proving Proposition 1

Lemma 10. *Let P_1, Q_1 be two root leaf paths in a network N_1 and let P_2 and Q_2 be two root-leaf paths respectively isomorphic to P_1 and Q_1, in another network N_2. Suppose that N_2 satisfies the NELP property and that, for every wishbone or crack K_1 contained in $P_1 \cup Q_1$ and containing P_1, there exists in N_2 a wishbone or crack K_2 isomorphic to K_1. Then P_2 and Q_2 intersect each other at the same depths as P_1 and Q_1, that is:*

P_1 and Q_1 have a node u in common that is at depth δ in P_1 and depth δ' in Q_1 if and only if P_2 and Q_2 have a node v in common that is at depth δ in P_2 and depth δ' in Q_2.

Proof. In the following, we prove that if P_1 and Q_1 have in common the nodes $u_0, u_1, \ldots, u_k$ and only these nodes (where u_0 is the root of N_1), with respective depths $\delta_0 = 0 < \delta_1 < \ldots < \delta_k$ in P_1 and depths $\delta'_0 = 0 < \delta'_1 < \ldots < \delta'_k$ in Q_1, then in N_2 there must be exactly $k+1$ nodes in common between P_2 and Q_2, and these nodes must be at depths $\delta_0, \delta_1, \ldots, \delta_k$ in P_2 and depths $\delta'_0, \delta'_1, \ldots, \delta'_k$ in Q_2. This statement is clearly equivalent to the statement that we wish to prove.

First, let $v_0, v_1, \ldots, v_k$ be the nodes that have respective depths $\delta_0, \delta_1, \ldots, \delta_k$ in P_2. The existence of these nodes is guaranteed by the fact that P_2 is isomorphic to P_1, and P_1 has nodes $u_0, u_1, \ldots, u_k$ at exactly those depths (Lemma 3). Furthermore let Δ_P be the length of P_1 and P_2, let Δ_Q be the length of Q_1 and Q_2, let x_P be the taxon labelling the leaves of P_1 and P_2 and let x_Q be the taxon labelling the leaves of Q_1 and Q_2. Below, we prove the following two claims:

(C1) For every $i \in \{1, 2, \ldots, k\}$, there exists in N_2 a weighted path from v_{i-1} to v_i of length $\delta'_i - \delta'_{i-1}$ that has no node in common with P_2 other than v_{i-1} and v_i.

(C2) If v_k is a leaf, then it is labelled by x_Q; otherwise there exists in N_2 a weighted path from v_k to the leaf labelled by x_Q that has no node in common with P_2 other than v_k and whose length is equal to $\Delta_Q - \delta'_k$.

We start with the proof of C1. See Fig. S2 in order to follow the reasoning below. First consider the trivial case where (u_{i-1}, u_i) is an edge in both P_1 and Q_1, and that it is assigned the same length λ in both of them. In this case, it is clear that $\delta_i - \delta_{i-1} = \delta'_i - \delta'_{i-1} = \lambda$. Because P_2 is isomorphic to P_1 and because u_{i-1} and u_i are consecutive nodes in P_1, the nodes at the same depths as u_{i-1} and u_i in P_2, that is v_{i-1} and v_i, must also be consecutive in P_2 (by Lemma 3). That is, (v_{i-1}, v_i) is an edge of P_2. Its length in P_2 must be equal to the difference between the depths of v_{i-1} and v_i in P_2, that is $\delta_i - \delta_{i-1} = \delta'_i - \delta'_{i-1}$. Edge (v_{i-1}, v_i) along with its length in P_2 constitutes a weighted path from v_{i-1} to v_i of length $\delta'_i - \delta'_{i-1}$ that has no node in common with P_2 other than v_{i-1} and v_i, thus proving C1 in this trivial case. In all other cases, define R_1^i as the root-leaf path that shares with P_1 its prefix down to u_{i-1} and its suffix from u_i onwards, and shares with Q_1 its portion between u_{i-1} and u_i. By construction, Q_1 has no node in common with P_1 between u_{i-1} and u_i, meaning that R_1^i and P_1 have in common only a prefix and a suffix, and thus that $P_1 \cup R_1^i$ is a crack. Because $P_1 \cup R_1^i$ is a crack containing P_1 and contained in $P_1 \cup Q_1$, then there exists in N_2 a crack isomorphic to $P_1 \cup R_1^i$. Because isomorphic cracks are the union of isomorphic root-leaf paths (Lemma 5) and because in N_2 there can be no root-leaf path isomorphic to P_1 other than P_2 (as N_2 satisfies the NELP property), then the crack isomorphic to $P_1 \cup R_1^i$ in N_2 can be written as $P_2 \cup R_2^i$, where R_2^i is a root-leaf path isomorphic to R_1^i. Now note that P_1 and R_1^i have lengths Δ_P and $\Delta_P - (\delta_i - \delta_{i-1}) + (\delta'_i - \delta'_{i-1})$, respectively, and longest common prefix and suffix of lengths δ_{i-1} and $\Delta_P - \delta_i$, respectively (as δ_{i-1} and δ_i are the respective depths of u_{i-1} and u_i in P_1). By Lemma 5, also P_2 and R_2^i and their longest common prefix and suffix must have these lengths. This implies that R_2^i separates from P_2 at the node at depth δ_{i-1} in P_2 — which by construction is v_{i-1} — then follows a weighted path of length $\delta'_i - \delta'_{i-1}$ that has no node in common with P_2 other than its extremes, and finally joins up with P_2 at the node at depth δ_i in P_2 — which by construction is v_i. The portion of R_2^i between v_{i-1} and v_i has length $\delta'_i - \delta'_{i-1}$ and no node in common with P_2 other than v_{i-1} and v_i, thus proving claim C1.

As for C2, if v_k is a leaf, since v_k belongs to both P_2 and Q_2, then it is the leaf of both P_2 and Q_2, and thus it must be labelled by $x_P = x_Q$. If instead v_k is not a leaf, then it must have strict descendants in P_2. Then, because P_1 and P_2 are isomorphic, also u_k (the node at the same depth in P_1 as v_k in P_2) must have strict descendants and is thus not a leaf. Define then R_1^{k+1} as the root-leaf path that shares with P_1 its prefix down to u_k, and shares with Q_1 its suffix from u_k to the leaf labelled by x_Q (as u_k is not a leaf, this suffix contains at least one edge). By construction, Q_1 has no node in common with P_1 after separating from it in u_k, meaning that R_1^{k+1} and P_1 have in common only a prefix, and thus that $P_1 \cup R_1^{k+1}$ is a wishbone. Because $P_1 \cup R_1^{k+1}$ is a wishbone containing P_1 and contained in $P_1 \cup Q_1$, then there exists in N_2 a wishbone isomorphic to $P_1 \cup R_1^{k+1}$. Because isomorphic wishbones are the union of isomorphic root-leaf paths (Lemma 4) and because in N_2 there can be no root-leaf path isomorphic to P_1 other than P_2 (as N_2 satisfies the NELP property), then the crack isomorphic to $P_1 \cup R_1^{k+1}$ in N_2 can be written as $P_2 \cup R_2^{k+1}$, where R_2^{k+1} is a root-leaf path isomorphic to R_1^{k+1}. Now note that P_1 and R_1^{k+1} have lengths Δ_P and $\Delta_Q - \delta'_k + \delta_k$, respectively, and longest common prefix of length δ_k. By Lemma 4, also P_2 and R_2^{k+1} and their longest common prefix must have these lengths. This implies that R_2^{k+1} separates from P_2 at the node at depth δ_k in P_2 — which by construction is v_k — and then follows a weighted path of length $\Delta_Q - \delta'_k$, that has no node in common with P_2 other than v_k, and that ends up in a leaf labelled by x_Q (as R_2^{k+1} is isomorphic to R_1^{k+1}). Thus C2 is also proved.

As a consequence of C1 and C2, one can construct a root-leaf path R in N_2 by concatenating all the weighted paths whose existence has been proven in C1 and C2. Clearly, R is a root-leaf path to x_Q and it has a total length of Δ_Q, as:

$$\begin{cases} \sum_{i=1}^{k} (\delta'_i - \delta'_{i-1}) = \delta'_k = \Delta_Q & \text{if } v_k \text{ is a leaf,} \\ \Delta_Q - \delta'_k + \sum_{i=1}^{k} (\delta'_i - \delta'_{i-1}) = \Delta_Q & \text{otherwise.} \end{cases}$$

Because in N_2 there can be no root-leaf path to x_Q of length Δ_Q other than Q_2 (as N_2 satisfies the NELP property), then $R = Q_2$. Now note that C1 and C2 imply that the only nodes that $R = Q_2$ has in common with P_2 are $v_0, v_1, \ldots, v_k$. Moreover, for every $i \in \{0, 1, \ldots, k\}$, the depth of v_i in P_2 equals δ_i (by definition), and the depth of v_i in $R = Q_2$ equals $\sum_{j=1}^{i}(\delta'_j - \delta'_{j-1}) = \delta'_i$, which is what we set out to prove. □

Proposition 1. *Two networks N_1 and N_2 with the NELP property are isomorphic if and only if they contain the same wishbones and cracks (up to isomorphism).*

Proof. The *only if* part is trivial: if N_1 and N_2 are isomorphic, than each wishbone or crack W contained in one of the two networks must have an isomorphic sub-network W' in the other (by Lemma 2), and W' must be either be a wishbone or a crack (by Lemmas 4 and 5).

As for the *if* part, let us now suppose networks $N_1 = (V_1, E_1, \varphi_1, \Lambda_1)$ and $N_2 = (V_2, E_2, \varphi_2, \Lambda_2)$ satisfy the NELP property and contain isomorphic wishbones and cracks. We prove that N_1 and N_2 are isomorphic.

Because N_1 and N_2 contain the same wishbones and cracks (up to isomorphism), and because a root-leaf path (which is both a wishbone and a crack) can only be isomorphic to another root-leaf path (Lemma 3), then N_1 and N_2 must contain the same root-leaf paths (up to isomorphism). Note that N_1 and N_2 must be networks on the same taxon set $\mathcal{X}$, because otherwise any root-leaf path that leads to a leaf corresponding to a taxon present in only one of the networks would have no isomorphic root-leaf path in the other network. Then, define a relation $\sim$ between V_1 and V_2 as follows: for any $v_1 \in V_1$ and $v_2 \in V_2$ we write $v_1 \sim v_2$ if there exist two isomorphic root-leaf paths P_1 and P_2, in N_1 and N_2, respectively, that contain v_1 and v_2, respectively, such that v_1 has the same depth in P_1 as v_2 in P_2. Note that, because N_1 and N_2 contain isomorphic root-leaf paths, for every $v_1 \in V_1$ there exists a $v_2 \in V_2$ such that $v_1 \sim v_2$ (by Lemma 3). Moreover, we now prove that such v_2 is unique. Suppose there exist v_2 and v'_2 such that $v_1 \sim v_2$ and $v_1 \sim v'_2$. This would mean (see Fig. S3) that there exist two isomorphic root-leaf paths P_1 and P_2, in N_1 and N_2, respectively, that have v_1 and v_2 at the same depth δ and that there exist two isomorphic root-leaf paths Q_1 and Q_2, in N_1 and N_2, respectively, that have v_1 and v'_2 at the same depth δ'. Because N_1 and N_2 contain the same wishbones and cracks, in particular for every wishbone or crack contained in $P_1 \cup Q_1$ and containing P_1 there exists an isomorphic wishbone or crack in N_2. Thus the assumptions of Lemma 10 are verified, meaning that P_2 and Q_2 must intersect each other at the same depths as P_1 and Q_1. As a result, because P_1 and Q_1 have a node v_1 in common that is at depth δ in P_1 and depth δ' in Q_1, then P_2 and Q_2 must have a node in common that is at depth δ in P_2 and depth δ' in Q_2. But v_2 and v'_2 are precisely the nodes in P_2 and Q_2 at depths δ and δ', respectively, meaning that we must have $v_2 = v'_2$. We thus conclude that there is a unique $v_2 \in V_2$ such that $v_1 \sim v_2$. Similarly, for every $v_2 \in V_2$ there exists a unique $v_1 \in V_1$ such that $v_1 \sim v_2$. Thus, relation $\sim$ identifies a bijection $f : V_1 \to V_2$, defined by $f(v_1) = v_2 \Leftrightarrow v_1 \sim v_2$.

We now prove that bijection f is an isomorphism between N_1 and N_2, by showing that it verifies the two requirements in Definition 5. First, note that for every $x \in \mathcal{X}$ we can consider a root-leaf path P_1 in N_1 ending in $\varphi_1(x)$ and its isomorphic equivalent P_2 in N_2 ending in $\varphi_2(x)$. Because the two paths are isomorphic, $\varphi_1(x)$ must lie at the same depth in P_1 as $\varphi_2(x)$ in P_2. Thus $\varphi_1(x) \sim \varphi_2(x)$, that is $f(\varphi_1(x)) = \varphi_2(x)$.

Second, we show that if $e_1 = (u, v) \in E_1$ is an edge of N_1 having a length $\lambda \in \Lambda_1(e_1)$ then $(f(u), f(v))$ is an edge of N_2 having also length λ. Let P_1 be any root-leaf path that passes via e_1 and assigns length λ to e_1. Let δ and $\delta + \lambda$ be the depths of u and v, respectively, in P_1. Because N_1 and N_2 contain isomorphic root-leaf paths, N_2 must contain a root-leaf path P_2, isomorphic to P_1. As a consequence of Lemma 3, P_2 must have two nodes at depths δ and $\delta + \lambda$ — which by construction must be $f(u)$ and $f(v)$, respectively — connected by an edge $e_2 = (f(u), f(v)) \in E_2$ having length λ in P_2, that is, $\lambda \in \Lambda_2(e_2)$. This allows to conclude that $\Lambda_1((u, v)) \subseteq \Lambda_2((f(u), f(v)))$. Similarly, we can prove that $\Lambda_2((f(u), f(v))) \subseteq \Lambda_1((u, v))$, which allows us to conclude that f satisfies point (ii) in Definition 5. Bijection f is thus an isomorphism

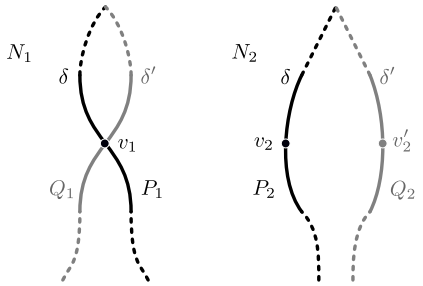


Figure S3. Illustration for the proof of Proposition 1.

between N_1 and N_2. □

Proving Proposition 2

In order to prove that two indistinguishable funnel-free networks with the NELP property have isomorphic wishbones and cracks, we need some more accessory results and notation.

Definition 9. Let P and Q be two root-leaf paths in a network N. We denote the length of the longest common prefix of P and Q by $\mu(P, Q)$. We say that P and Q *separate* at node v, if $P \neq Q$, and v is the last node in the longest common prefix of P and Q. Finally, P and Q *separate openly* at v if they separate at v and the direct descendants of v in P and Q are distinct nodes.

Note that if the longest common prefix of P and Q consists of only one node, then $\mu(P, Q) = 0$. In general, if P and Q separate at v, then the depth of v in P and Q is precisely $\mu(P, Q)$. Moreover, note that two distinct root-leaf paths P and Q do not separate openly at v when they contain the same edge (v, v'), but with different lengths in P and Q.

Definition 10. Let P and Q denote root-leaf paths. An *open crack* is a crack $P \cup Q$ where P and Q separate openly. A *closed crack* is a crack $P \cup Q$ where P and Q differ for the length of exactly one edge.

Note that all cracks are either open, closed or root-leaf paths. In Fig. 14 in the main text, S and R are an open crack and a closed crack, respectively.

Lemma 11. *(Three-prefix condition). Let P_1, P_2, P_3 be root-leaf paths in a network N. Then the two smallest of $\mu(P_1, P_2)$, $\mu(P_1, P_3)$, $\mu(P_2, P_3)$ are equal, or, equivalently, for $\{i, j, k\} = \{1, 2, 3\}$:*

$$\mu(P_i, P_j) \geq \min\{\mu(P_i, P_k), \mu(P_j, P_k)\}.$$

Proof. First note that the two smallest of three numbers x_1, x_2, x_3 are equal if and only if $x_i \geq \min\{x_j, x_k\}$, for $\{i, j, k\} = \{1, 2, 3\}$: without loss of generality suppose $x_1 \leq x_2 \leq x_3$ and note that — while $x_2 \geq \min\{x_1, x_3\} = x_1$ and $x_3 \geq \min\{x_1, x_2\} = x_1$ are trivially true — $x_1 \geq \min\{x_2, x_3\}$ holds if and only if $x_1 = x_2$. In order to prove that the two smallest of $\mu(P_1, P_2)$, $\mu(P_1, P_3)$, $\mu(P_2, P_3)$ are equal, imagine following P_1, P_2 and P_3 from the root until at least one of them separates from the others: let v be a node at depth δ in P_1, P_2, P_3, where either only P_i separates from P_j and P_k, or all three P_1, P_2, P_3 separate. In both cases, $\delta = \mu(P_i, P_j) = \mu(P_i, P_k) \leq \mu(P_j, P_k)$, thus concluding the proof of this lemma. □

Lemma 12. *Let N and N' be two indistinguishable funnel-free networks, and let N' satisfy the NELP property. Let P_1 and P_2 be root-leaf paths in N, whose union $P_1 \cup P_2$ is a wishbone. Let P_1' and P_2'*

be the root-leaf paths in N' to the same taxa as P_1 and P_2, and having the same lengths as P_1 and P_2, respectively (whose existence and uniqueness are guaranteed by Corollary 3 and Lemma 9, respectively). Then, $P'_1 \cup P'_2$ is a wishbone of N', and

$$\mu(P'_1, P'_2) = \mu(P_1, P_2).$$

Proof. Let x_1 and x_2 be the taxa labelling the leaves of P_1 and P_2 (and thus P'_1 and P'_2), respectively, and let λ_1 and λ_2 be the lengths of P_1 and P_2 (and thus P'_1 and P'_2). Note that we may have $x_1 = x_2$, and $\lambda_1 = \lambda_2$, if $P_1 = P_2$. Because N and N' are indistinguishable, they must both weakly display T_W, the tree obtained by suppressing all suppressible nodes in the wishbone $W = P_1 \cup P_2$ (by Proposition 3). T_W must be the union of two root-leaf paths to x_1 and x_2, of lengths λ_1 and λ_2, respectively, and having a longest common prefix of length $\mu(P_1, P_2)$ (because these properties are not lost by suppressing suppressible nodes). Then, the embedding of T_W in N' must be a wishbone W' consisting of the union of two root-leaf paths to x_1 and x_2, of lengths λ_1 and λ_2, respectively, whose longest common prefix has length $\mu(P_1, P_2)$ (by Lemma 6). Because P'_1 and P'_2 are the unique root-leaf paths in N' to x_1 and x_2, and of lengths λ_1 and λ_2, respectively, this implies $W' = P'_1 \cup P'_2$ and the lemma follows. □

Note that the lemma that we just proved includes the trivial case where $P_1 = P_2$.

Lemma 13. *Let N and N' be two indistinguishable funnel-free networks, and let N' satisfy the NELP property. Let P and Q be distinct root-leaf paths in N, whose union $P \cup Q$ is a crack. Let P' and Q' be the root-leaf paths in N' to the same taxa, and having the same lengths as P and Q, respectively (whose existence and uniqueness are guaranteed by Corollary 3 and Lemma 9, respectively). Then*

$$\mu(P', Q') \begin{cases} = \mu(P, Q) & \text{if } P \cup Q \text{ is an open crack,} \\ \geq \mu(P, Q) & \text{if } P \cup Q \text{ is a closed crack.} \end{cases}$$

Proof. The proof is by induction on the number of edges in the longest suffix common to P and Q. Throughout this proof, for any root-leaf path X in N, let X' denote the root-leaf path in N' to the same taxon and having the same length as X. Moreover, let u be the last node in the longest prefix common to P and Q and let v be the first node in the longest suffix common to P and Q. Because $P \neq Q$, u must be a strict ancestor of v. Moreover, let u_P and u_Q be the direct ancestors of v in P and Q, respectively. Of these two nodes, at least one is not a strict ancestor of the other, otherwise N would contain a cycle. Thus, without loss of generality, we assume throughout that u_Q is not a strict ancestor of u_P. Note that if $P \cup Q$ is an open crack, at least one between u_P and u_Q must be different from u. Because we require that u_Q is not a strict ancestor of u_P, it follows that $u_Q \neq u$, when $P \cup Q$ is an open crack. Finally, in the particular case where $P \cup Q$ is a closed crack and $u = u_Q$ is the root of N, the statement trivially holds, as $\mu(P, Q) = 0 \leq \mu(P', Q')$. Thus, we assume throughout that u_Q is not the root of N. Because it cannot be a leaf or a funnel, either, we can assume that u_Q has outdegree 2 or more.

Base case. (See Fig. S4, left, to follow the argument below.) Suppose that the longest suffix common to P and Q contains no edge, or, equivalently, that it consists of just a leaf v. Because u_Q has outdegree 2 or more, we can define a root-leaf path R in N that separates openly from Q at u_Q: let R have a common prefix with Q consisting of the portion of Q from the root down to u_Q; then, let R take an edge (u_Q, w) with $w \neq v$ (with any of this edge's lengths in N) and finally let R take any weighted path from w that does not end up in v (which is possible because N is funnel-free). Clearly,

$$\mu(P, Q) = \mu(P, R) \leq \mu(Q, R). \tag{3}$$

Note that R and Q have no node in common below u_Q — as otherwise R would contain v, which we excluded by construction — meaning that $Q \cup R$ is a wishbone. Moreover, $P \cup R$ is also a wishbone: assuming otherwise would mean that R has a node in common with P below u, either contradicting the

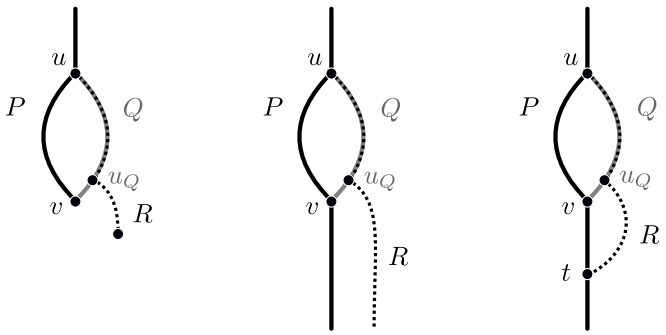


Figure S4. Illustration of the base case (left) and the inductive step (other two drawings) in the proof of Lemma 13. P is shown in black, the portion of Q not coinciding with P is shown in grey, and the portion of R not coinciding with P is represented by the dashed line. Note that, although the drawing shows the case of an open crack with $u \neq u_Q$, the proof also considers the case where $u = u_Q$, i.e. when $P \cup Q$ is a closed crack.

fact that P and Q have no nodes in common between u and v, or contradicting the requirement that u_Q is not a strict ancestor of u_P, or contradicting the requirement that v does not belong to R. Because $Q \cup R$ and $P \cup R$ are wishbones, then (by Lemma 12)

$$\mu(Q', R') = \mu(Q, R), \qquad \mu(P', R') = \mu(P, R) \tag{4}$$

Now consider two cases: either $P \cup Q$ is an open crack or it is closed. If it is open, then $u \neq u_Q$ (see above) and thus we have $\mu(P, R) < \mu(Q, R)$ in (3). Then, because of the equalities in (4), we have $\mu(P', R') < \mu(Q', R')$. But then, because of the three-prefix condition (Lemma 11), we must have $\mu(P', Q') = \mu(P', R')$. Combine this with Equations (3) and (4), to show that $\mu(P', Q') = \mu(P, Q)$. If instead $P \cup Q$ is a closed crack, then $u = u_Q$ and thus we have $\mu(P, R) = \mu(Q, R)$. Then, because of the equalities in (4), we have $\mu(P', R') = \mu(Q', R')$. In this case, the three-prefix condition (Lemma 11) implies $\mu(P', Q') \geq \mu(P', R') = \mu(Q', R')$ and thus $\mu(P', Q') \geq \mu(P, Q)$.

Inductive step. (See the two drawings on the right in Fig. S4 to follow the argument below.) Now suppose that the longest suffix common to P and Q contains $k > 0$ edges, and assume that the lemma's statements hold for every pair of root-leaf paths whose union is a crack and whose longest common suffix contains fewer than k edges. Because u_Q has outdegree 2 or more, we can define a root-leaf path R in N that separates openly from Q at u_Q: let R have a common prefix with Q consisting of the prefix of Q from the root down to u_Q; then, let R take an edge (u_Q, w) with $w \neq v$ (with any of this edge's lengths in N) and then let R continue taking edges always avoiding ending up in v (recall that N is funnel-free), until it either arrives in a leaf or in a node t belonging to $P \cup Q$. Note that such t must be a strict descendant of v in the suffix common to P and Q. (By construction $t \neq v$; moreover, if t belonged to Q and not to P, then t would lie between u_Q and v in Q, which contradicts the assumption that u_Q is a direct ancestor of v in Q; finally, if t belonged to P and not to Q, then t would either coincide with u_P or be a strict ancestor of u_P, implying that u_Q is an ancestor of u_P, which is not possible by construction.) Finally, from t onwards, let R coincide with the suffix common to P and Q. Clearly, as in the proof of the base case, the following holds:

$$\mu(P, Q) = \mu(P, R) \leq \mu(Q, R). \tag{5}$$

If R arrives in a leaf without hitting a node in $P \cup Q$ (i.e., without crossing neither P or Q), then $Q \cup R$ and $P \cup R$ are wishbones. If instead R hits $P \cup Q$ in t, then $Q \cup R$ and $P \cup R$ are open cracks: they are cracks, because, by construction, t is the first node that R has in common with P and Q after separating

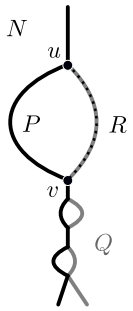


Figure S5. Illustration of the inductive case in the proof of Lemma 14. P is shown in black, the portion of Q not coinciding with P is shown in grey, and the portion of R not coinciding with P is represented by the dashed line.

from them (at u and u_Q, respectively), and because R has its suffix following t in common with P and Q; they are open because the edge (u_Q, w) in R does not belong to neither Q or P. Moreover, these open cracks are such that the two root-leaf paths that compose them have a longest common suffix of fewer than k edges, as t is a strict descendant of v. Irrespective of $Q \cup R$ and $P \cup R$ being wishbones or open cracks, we then have (by Lemma 12 or by inductive hypothesis) that

$$\mu(Q', R') = \mu(Q, R), \qquad \mu(P', R') = \mu(P, R). \tag{6}$$

Now consider two cases: either $P \cup Q$ is an open crack or it is closed. If it is open, then $u \neq u_Q$ (see above) and thus we have $\mu(P, R) < \mu(Q, R)$ in (5). Then, because of the equalities in (6), we have $\mu(P', R') < \mu(Q', R')$. But then, because of the three-prefix condition (Lemma 11), we must have $\mu(P', Q') = \mu(P', R')$. If we combine this with Equations (5) and (6), then we deduce that $\mu(P', Q') = \mu(P, Q)$. If instead $P \cup Q$ is a closed crack, then $u = u_Q$ and thus we have $\mu(P, R) = \mu(Q, R)$. Then, because of the equalities in (6), we have $\mu(P', R') = \mu(Q', R')$. In this case, the three-prefix condition (Lemma 11) implies $\mu(P', Q') \geq \mu(P', R') = \mu(Q', R')$ and thus $\mu(P', Q') \geq \mu(P, Q)$. □

Lemma 14. *Let N and N' be two indistinguishable funnel-free networks, and let N' satisfy the NELP property. Let P and Q be any root-leaf paths in N, and let P' and Q' be the root-leaf paths in N' to the same taxa, and having the same lengths as P and Q, respectively (whose existence and uniqueness are guaranteed by Corollary 3 and Lemma 9, respectively). Then, (i)*

$$\mu(P', Q') \begin{cases} = \mu(P, Q) & \text{if } P = Q \text{ or if they separate openly,} \\ \geq \mu(P, Q) & \text{otherwise.} \end{cases}$$

Moreover, (ii) if P and Q separate openly, then P' and Q' also separate openly.

Proof. Once again, throughout this proof, for any root-leaf path X in N, we let X' denote the root-leaf path in N' to the same taxon and having the same length as X. Let $\Delta(P, Q)$ denote the number of edges in $P \cup Q$ that are either present in only one of P and Q, or that are present in both, but with different lengths. We prove part (i) by induction on $\Delta(P, Q)$.

Base case. If $\Delta(P, Q) = 0$, then $P = Q$ and $P' = Q'$, trivially implying that both $\mu(P, Q)$ and $\mu(P', Q')$ equal the length of these paths, and thus $\mu(P', Q') = \mu(P, Q)$.

Inductive step. If $\Delta(P, Q) > 0$, then P and Q must separate at a node u. If they separate openly and do not have any node in common after u, then they form a wishbone, meaning that $\mu(P', Q') = \mu(P, Q)$ (by Lemma 12), thus verifying the statement. In all other cases, let v be the first node that P and Q have

in common after u. Define a new root-leaf path R that coincides with P and Q along all their common prefix down to u, then coincides with Q along the weighted path in Q from u to v, and finally coincides with P along all its suffix from v to the leaf in P (see Fig. S5). Clearly,

$$\mu(P,Q) = \mu(P,R) < \mu(Q,R). \tag{7}$$

Note that $\Delta(Q,R) = \Delta(P,Q) - \Delta(P,R)$. Because $P \neq R$, then $\Delta(P,R) > 0$. Therefore, we have $\Delta(Q,R) < \Delta(P,Q)$ and we can assume, by inductive hypothesis:

$$\mu(Q',R') \geq \mu(Q,R). \tag{8}$$

Because P and R are node-disjoint between u and v, and coincide everywhere else, clearly they form a crack. This crack is open or closed, depending on whether P and Q separate openly or by taking the same edge (u,v) with different lengths. Consider these two cases separately.

If P and Q separate openly, then $P \cup R$ is an open crack and Lemma 13 implies that

$$\mu(P',R') = \mu(P,R). \tag{9}$$

Now combine Equations (7), (8), (9) to show that $\mu(P',R') < \mu(Q',R')$. But then, because of the three-prefix condition (Lemma 11), we must have $\mu(P',Q') = \mu(P',R')$, which together with Equations (7) and (9) implies $\mu(P',Q') = \mu(P,Q)$, when P and Q separate openly.

If instead P and Q separate by taking the same edge (u,v) with different lengths, then $P \cup R$ is a closed crack. Then, by Lemma 13,

$$\mu(P',R') \geq \mu(P,R). \tag{10}$$

Now combine Equations (7), (8), (10) to show that $\mu(P,Q) \leq \mu(Q',R')$ and $\mu(P,Q) \leq \mu(P',R')$. But then, because of the three-prefix condition (Lemma 11), we must have

$$\mu(P',Q') \geq \min\{\mu(Q',R'), \mu(P',R')\} \geq \mu(P,Q),$$

which concludes our proof by induction of part (i).

As for part (ii), suppose that P and Q separate openly at a node u, at depth α in their common prefix, by taking two distinct edges (u,w_P) and (u,w_Q). Because N is funnel-free, because w_P and w_Q are distinct, and because neither of them is the root of N, then there exist two node-disjoint directed paths π_P and π_Q from w_P and w_Q, respectively, to two leaves of N (by Proposition 4). Let R_P be any root-leaf path that coincides with P along its prefix down to w_P and then follows π_P by taking its edges with any of their lengths in N. Similarly, let R_Q be any root-leaf path that coincides with Q along its prefix down to w_Q and then follows π_Q by taking its edges with any of their lengths in N. Because R_P and R_Q coincide with P and Q down to node u, and all their nodes that are strict descendants of u belong to (the node-disjoint paths) π_P and π_Q, respectively, then $R_P \cup R_Q$ is a wishbone in N, with $\mu(R_P,R_Q) = \alpha$. By Lemma 12, also $R'_P \cup R'_Q$ is a wishbone in N' and $\mu(R'_P,R'_Q) = \alpha$.

Moreover, because, by construction, $\mu(P,R_P) > \alpha$ and $\mu(Q,R_Q) > \alpha$, and because, by part (i) of the present lemma, $\mu(P',R'_P) \geq \mu(P,R_P)$ and $\mu(Q',R'_Q) \geq \mu(Q,R_Q)$, then we have

$$\mu(P',R'_P) > \alpha \qquad \text{and} \qquad \mu(Q',R'_Q) > \alpha.$$

These two relationships, together with the fact that R'_P and R'_Q separate openly at depth α, imply that P' and Q' also separate openly at depth α: if we let u' be the node at depth α in R'_P and R'_Q, the successors of u' in R'_P and R'_Q must be distinct, and belonging to P' and Q', respectively, thus implying that P' and Q' separate openly. □

The lemmas above only require one of the two networks to verify the NELP property. In the rest of this section, we concentrate on the case where both networks satisfy the NELP property, as this is among the hypotheses of Proposition 2.

Lemma 15. *Let N and N' be two indistinguishable funnel-free networks satisfying the NELP property. Let P and Q be any root-leaf paths in N, and let P' and Q' be the root-leaf paths in N' to the same taxa, and having the same lengths as P and Q, respectively (whose existence and uniqueness are guaranteed by Corollary 3 and Lemma 9, respectively). Then, P' and Q' separate openly if and only if P and Q separate openly, and*

$$\mu(P', Q') = \mu(P, Q).$$

Proof. Apply Lemma 14 both to P and Q in N and to P' and Q' in N', showing that $\mu(P', Q') \geq \mu(P, Q)$ and $\mu(P, Q) \geq \mu(P', Q')$, respectively (and thus $\mu(P', Q') = \mu(P, Q)$) and that if one pair of root-leaf paths separates openly, also the other separates openly. □

Lemma 16. *Let N and N' be two indistinguishable funnel-free networks satisfying the NELP property. Then N and N' have the same root-leaf paths (up to isomorphism). Moreover, for each root-leaf path P in one of the two networks, the root-leaf path P' isomorphic to P in the other network is unique.*

Proof. Let P be a root-leaf path in one of the two networks, say (without loss of generality), N, and let P' be the unique root-leaf path in N' to the same taxon and having the same length as P (whose existence and uniqueness are guaranteed by Corollary 3 and Lemma 9, respectively). For any node v in P, there exists a node v' at the same depth in P': this is trivial when v is at depth 0; otherwise, if we let Q be a root-leaf path that separates openly from P at v (which exists because N is funnel-free), then Q' — the root-leaf path in N' to the same taxon and having the same length as Q — must separate from P' at a node at the same depth $\mu(P', Q') = \mu(P, Q)$ as v (by Lemma 14). Symmetrically, for any node in P' there exists a node at the same depth in P. Thus, because P and P' are root-leaf paths to the same taxon and have nodes at the same depths, then P and P' are isomorphic (by Lemma 3). In conclusion, for any root-leaf path P in one of the two networks, there exists a unique isomorphic root-leaf path P' in the other, which is what we wanted to prove. □

Lemma 17. *Let N and N' be two indistinguishable funnel-free networks satisfying the NELP property. Let P_1 and P_2 be root-leaf paths in N, whose union $W = P_1 \cup P_2$ is a wishbone. Let P_1' and P_2' be the unique root-leaf paths in N' isomorphic to P_1 and P_2, respectively (Lemma 16). Then $W' = P_1' \cup P_2'$ is a wishbone in N' isomorphic to W.*

Proof. By Lemma 12, $W' = P_1' \cup P_2'$ is a wishbone with $\mu(P_1', P_2') = \mu(P_1, P_2)$. Because W and W' are wishbones, and equal to the union of pairs of isomorphic root-leaf paths, with longest common prefixes of the same length, then W and W' are isomorphic (by Lemma 4). □

Corollary 4. *Let N and N' be two indistinguishable funnel-free networks satisfying the NELP property. Then N and N' have the same wishbones (up to isomorphism).*

Proof. Apply Lemma 17 to all wishbones in N and N'. □

Lemma 18. *Let N and N' be two indistinguishable funnel-free networks with the NELP property. Then they have the same cracks (up to isomorphism).*

Proof. We prove that for any crack $K = P_1 \cup P_2$ contained in one of the two networks, there exists a crack K', isomorphic to K, in the other network. Without loss of generality, we assume here that K is a crack in N, but all arguments hold symmetrically for the case where K is a crack in N'.

We introduce some notation that is useful throughout the proof. For any root-leaf path X in N, let X' denote the unique root-leaf path in N' that is isomorphic to X (Lemma 16). The thesis is trivial when $P_1 = P_2$: in this case K is a root-leaf path, and K' is isomorphic to it. Thus we assume throughout that $P_1 \neq P_2$. Now let u be the last node in the longest prefix common to P_1 and P_2 and let v be the first node in the longest suffix common to P_1 and P_2. Because $P_1 \neq P_2$, u must be a strict ancestor of v. Also, let $\alpha = \mu(P_1, P_2) = \mu(P_1', P_2')$ (the equality is guaranteed by Lemma 15), meaning that P_1' and P_2'

separate at a node u' that has the same depth α in P'_1 and P'_2 as u in P_1 and P_2. Moreover let β_1 and β_2 be the (strictly positive) lengths of the weighted paths from u to v within P_1 and P_2, respectively, and let $\gamma \geq 0$ be the length of the longest suffix common to P_1 and P_2. Because v has depth $\alpha + \beta_1$ and $\alpha + \beta_2$ in P_1 and P_2, respectively, there must exist in N' two nodes v'_1 and v'_2 at depths $\alpha + \beta_1$ and $\alpha + \beta_2$ in P'_1 and P'_2, respectively. Finally, note that because P_1 and P_2 have a common suffix, they are root-leaf paths to the same taxon. Therefore also P'_1 and P'_2 are root-leaf paths to the same taxon, and thus end up in the same leaf. Thus P'_1 and P'_2 must share at least one node after separating at u'. Let v' be the first node that P'_1 and P'_2 have in common after separating at u' (thus a strict descendant of u'). All these notations are shown in Fig. S6.

We now prove that $v' = v'_1 = v'_2$ implies that $K' = P'_1 \cup P'_2$ is a crack isomorphic to K. First, note that the suffixes of P'_1 and P'_2 following v'_1 and v'_2, respectively, are weighted paths of length γ (because P'_1 and P'_2 have lengths $\alpha + \beta_1 + \gamma$ and $\alpha + \beta_2 + \gamma$, respectively), ending up in the same leaf. Thus, when $v'_1 = v'_2$, these two weighted paths have equal lengths and the same endpoints. Because N' has the NELP property, these two weighted paths must then coincide, meaning that P'_1 and P'_2 have a common suffix of length γ. Moreover, by the definition of v', P'_1 and P'_2 have no node in common between u' and v'. Therefore $v' = v'_1 = v'_2$ implies that $K' = P'_1 \cup P'_2$ is a crack with a longest common suffix of length γ, which is the same as the length of the longest common suffix of P_1 and P_2. Because of this, because P'_1 and P'_2 have a longest common prefix with the same length as the longest common prefix of P_1 and P_2 (α), and finally because P'_1 and P'_2 are isomorphic to P_1 and P_2, respectively, then K and K' are isomorphic (by Lemma 5).

Now consider the case where K is a closed crack: in this case, P_1 and P_2 only differ for the length assigned to edge (u, v), which must be β_1 in P_1 and β_2 in P_2. Note that P'_1 and P'_2 cannot separate openly, because otherwise also P_1 and P_2 would separate openly (by Lemma 15). Therefore, P'_1 and P'_2 must both contain the edge (u', v'), but assign different lengths to it. (We have called the head of this edge v', as it is clearly the first node that P'_1 and P'_2 have in common after separating at u'.) Because P'_1 is isomorphic to P_1, the length assigned to (u', v') in P'_1 must be β_1: if this edge were assigned a length $\beta'_1 < \beta_1$ in P'_1 , then a node at depth $\alpha + \beta'_1$ would exist in P'_1 and therefore in P_1, but this contradicts the fact that there is no node in P_1 between u (at depth α) and v (at depth $\alpha + \beta_1$); if instead (u', v') were assigned a length $\beta'_1 > \beta_1$ in P'_1, then no node would exist at depth $\alpha + \beta_1$ in P'_1, which contradicts the fact that v has depth $\alpha + \beta_1$ in P_1. Symmetrically, because P'_2 is isomorphic to P_2, the length assigned to (u', v') in P'_2 must be β_2. Thus v' has depth $\alpha + \beta_1$ in P'_1 and $\alpha + \beta_2$ in P'_2, meaning that $v' = v'_1 = v'_2$. As we showed above, this implies that $K' = P'_1 \cup P'_2$ is a crack isomorphic to K, whenever K is a closed crack.

It remains to prove that for any open crack $K = P_1 \cup P_2$ contained in one of the two networks, there exists a crack K' in the other network that is isomorphic to K. We prove this by induction on the number of edges in the longest suffix common to P_1 and P_2. In both the base case and the inductive step, we let u_1 and u_2 be the direct ancestors of v in P_1 and P_2, respectively. Of these two nodes, at least one is not an ancestor of the other, otherwise either N would contain a cycle or $u_1 = u_2$ (the latter would imply that K is a closed crack). Thus, without loss of generality, we assume throughout that u_2 is not an ancestor of u_1. Note that this implies that $u_2 \neq u$.

Base case. (See Fig. S7 to follow the argument below.) If the longest suffix common to P_1 and P_2 contains no edge, then v coincides with the leaf in P_1 and P_2, and $\gamma = 0$. Thus P'_1 and P'_2 have lengths $\alpha + \beta_1$ and $\alpha + \beta_2$, respectively, meaning that v'_1 and v'_2 are their leaves. But P'_1 and P'_2 have the same leaf, thus implying $v'_1 = v'_2$. It remains to prove that $v' = v'_1 = v'_2$, or in other words that P'_1 and P'_2 have no node in common between u' and $v'_1 = v'_2$. Because u_2 has outdegree 2 or more, we can define a root-leaf path R in N that separates openly from P_2 at u_2: let R have a common prefix with P_2 consisting of the portion of P_2 from the root down to u_2; then, let R take an edge (u_2, w) with $w \neq v$ (with any of this edge's lengths in N) and finally let R take any weighted path from w that does not end up in v (which is possible because N is funnel-free). Note that R and P_2 have no node in common below u_2 — as otherwise R would contain v, which we excluded by construction — meaning that $P_2 \cup R$ is a wishbone.

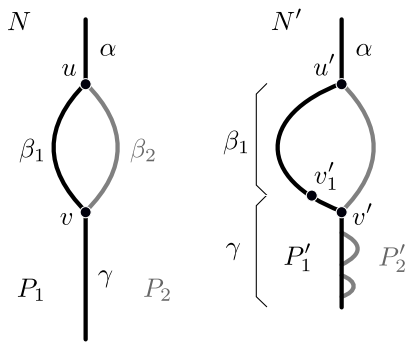


Figure S6. Illustration of the notation for Lemma 18 and of the argument against claim **C1** within it. P_1 and P'_1 are in black and the portions of P_2 and P'_2 not overlapping with P_1 and P'_1 are in grey. Note that the position of v'_2 along P'_2 is not shown as it may be either above or below v'.

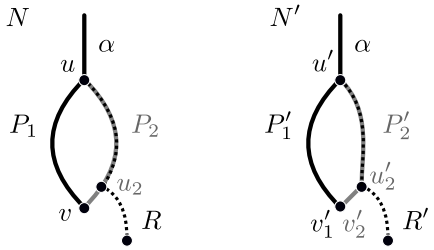


Figure S7. Illustration of the base case of the proof of Lemma 18. P_1 and P'_1 are in black and the portions of P_2 and P'_2 not overlapping with P_1 and P'_1 are in grey. Finally, the portions of R and R' not overlapping with P_1 and P'_1 are represented by the dashed lines.

Moreover, $P_1 \cup R$ is also a wishbone: assuming otherwise would mean that R has a node in common with P_1 below u, either contradicting the fact that P_1 and P_2 have no nodes in common between u and v, or contradicting the requirement that u_2 is not an ancestor of u_1, or contradicting the requirement that v does not belong to R. Now let $W_1 = P_1 \cup R$ and $W_2 = P_2 \cup R$. Because these are wishbones, by Lemma 17, $W'_1 = P'_1 \cup R'$ and $W'_2 = P'_2 \cup R'$ are also wishbones isomorphic to W_1 and W_2, respectively. Moreover, because the same holds for P_1, P_2 and R in N, the following holds:

$$\mu(P'_1, P'_2) = \mu(P'_1, R') = \alpha \leq \mu(P'_2, R'),$$

meaning that P'_1 and R' separate at the same node, u', where P'_1 and P'_2 separate.

Now recall that v' is the first node that P'_1 and P'_2 have in common after separating at u'. Note that v' cannot be in the prefix common to P'_2 and R', because otherwise v' would be a node common to P'_1 and R', which (together with the fact that v' is a strict descendant of u') contradicts the fact that P'_1 and R' form a wishbone and separate at u'. Thus v' must belong to the suffix of P'_2 after separation from R', or, in other words, v' must have a depth in P'_2 strictly greater than $\mu(P'_2, R')$. Now note that the only node in P_2 at a depth strictly greater than $\mu(P_2, R)$ is v, meaning (as P'_2 is isomorphic to P_2) that there can only be one node in P'_2 at a depth strictly greater than $\mu(P'_2, R') = \mu(P_2, R)$. This node is $v'_1 = v'_2$, thus allowing to conclude that $v' = v'_1 = v'_2$.

Inductive case. Now suppose that the longest suffix common to P_1 and P_2 contains $k > 0$ edges, and

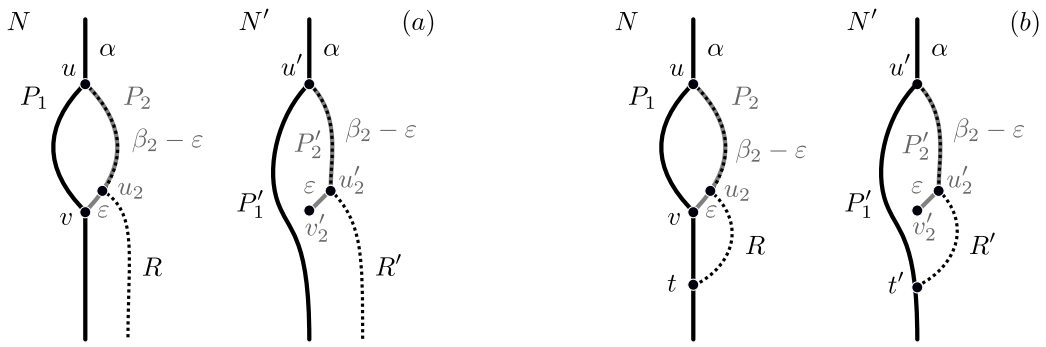


Figure S8. Illustration of the argument against claim **C3** in the proof of Lemma 18. P_1 and P'_1 are in black and the portions of P_2 and P'_2 not overlapping with P_1 and P'_1 are in grey. Finally, the portions of R and R' not overlapping with P_1 and P'_1 are represented by the dashed lines. In (a), P_1 and R — and thus P'_1 and R' — form a wishbone, whereas in (b), P_1 and R — and thus P'_1 and R' — form an open crack.

assume that for every open crack with up to $k-1$ edges in its longest common suffix in one of the two networks, there exists an isomorphic crack in the other network. We show that each of the following claims leads to contradiction: (**C1**) v' is a strict descendant of v'_1 along P'_1; (**C2**) v' is a strict descendant of v'_2 along P'_2; (**C3**) v' is a strict ancestor of v'_2 along P'_2; (**C4**) $v' = v'_2$ and v' is a strict ancestor of v'_1 along P'_1.

First, let us deal with claim **C1**: suppose that v' is a strict descendant of v'_1 along P'_1 (see Fig. S6). We show that for every wishbone or crack W' contained in $P'_1 \cup P'_2$ and containing P'_1, there exists in N a wishbone or crack W isomorphic to W'. This is trivial when W' is a wishbone: because P'_1 and P'_2 are root-leaf paths to the same taxon, the only wishbone W' contained in $P'_1 \cup P'_2$ and containing P'_1 is P'_1 itself, for which $W = P_1$. This is also trivial when W' is a closed crack, as we have already proved that N and N' must have the same closed cracks (up to isomorphism). It remains the case where W' is an open crack. In this case, W' can be obtained by combining P'_1 with a weighted path contained in P'_2 that has no node or edge in common with P'_1, other than its endpoints w and z. Because the first such weighted path in P'_2 is the one between $w = u'$ and $z = v'$, it follows that z must be a descendant of v' and thus a strict descendant of v'_1 along P'_1. Now note that, because P'_1 is isomorphic to P_1, and because the suffix of P_1 starting in v contains k edges, then the suffix of P'_1 starting in v'_1 (the node at the same depth in P'_1 as v in P_1) also contains exactly k edges. But then, because z is a strict descendant of v'_1 along P'_1, then the suffix of P'_1 starting in z contains strictly less than k edges. That is, the longest suffix common to the two root-leaf paths composing the open crack W' contains strictly less than k edges. Then, by inductive hypothesis, there must be a crack W in N that is isomorphic to W'. We have thus proved that for every wishbone or crack W' contained in $P'_1 \cup P'_2$ and containing P'_1, there exists in N a wishbone or crack W isomorphic to W'. This, together with the fact that N satisfies the NELP property, allows us to apply Lemma 10 and conclude that the unique root-leaf paths isomorphic to P'_1 and P'_2 in N, that is P_1 and P_2, must intersect each other at the same depths as P'_1 and P'_2. But this leads to a contradiction, as the node at depth $\alpha + \beta_1$ in P_1, that is v, belongs to both P_1 and P_2, whereas the node at the same depth in P'_1, that is v'_1 only belongs to P'_1. Similarly, one can prove that claim **C2** leads to contradiction.

Now assume (claim **C3**) that v' is a strict ancestor of v'_2 along P'_2. Recall that u_1 and u_2 are the two direct ancestors of v in P_1 and P_2, respectively, with u_2 assumed to not be an ancestor of u_1. Because u_2 has outdegree 2 or more, we can define a root-leaf path R in N that separates openly from P_2 at u_2: let R have a common prefix with P_2 consisting of the prefix of P_2 from the root down to u_2; then, let R take an edge (u_2, w) with $w \neq v$ (with any of this edge's lengths in N) and then let R continue taking

edges always avoiding ending up in v (recall that N is funnel-free), until it either arrives in a leaf or in a node t belonging to $P_1 \cup P_2$. Note that, if such t exists, then it must be a strict descendant of v in the suffix common to P_1 and P_2. (By construction $t \neq v$; moreover, if t belonged to P_2 and not to P_1, then t would lie between u_2 and v in P_2, which contradicts the assumption that u_2 is a direct ancestor of v in P_2; finally if t belonged to P_1 and not to P_2, then t would be an ancestor of u_1, implying that u_2 is an ancestor of u_1, which is not possible by construction.) Finally, from t onwards, let R coincide with the suffix common to P_1 and P_2. (See Fig. S8 to follow the argument below.)

Now consider $P_1 \cup R$. Because $P_1 \cup P_2$ is a crack and u_2 is strictly between u and v, R cannot have any node in common with P_1 between u and u_2 (u_2 included). Thus, if R arrives in a leaf without hitting a node in $P_1 \cup P_2$, then $P_1 \cup R$ is a wishbone. Then, by Lemma 17, $P'_1 \cup R'$ is a wishbone isomorphic to $P_1 \cup R$ (see Fig. S8(a)). If instead R hits $P_1 \cup P_2$ in t, then $P_1 \cup R$ is an open crack: it is a crack because, by construction, t is the first node that R has in common with P_1 after separating from it at u, and because R has its suffix following t in common with P_1; it is open, because $u_2 \neq u$ implies that the path from u to t in R is composed by more than one edge. Moreover the suffix of P_1 and R following t contains fewer than k edges, as t is a strict descendant of v. Thus, by inductive hypothesis, in N' there exists a crack isomorphic to $P_1 \cup R$. Because isomorphic cracks are the union of isomorphic root-leaf paths (Lemma 5) and because in N' there can be no root-leaf path isomorphic to P_1 and R other than P'_1 and R', respectively (as N' satisfies the NELP property, by Lemma 9), then this crack can be written as $P'_1 \cup R'$ (see Fig. S8(b)). Thus, we have proved that $P'_1 \cup R'$ is either a wishbone or a crack isomorphic to $P_1 \cup R$. Moreover, because the same holds for P_1, P_2 and R in N, the following holds:

$$\mu(P'_1, P'_2) = \mu(P'_1, R') = \alpha \leq \mu(P'_2, R') = \alpha + \beta_2 - \varepsilon,$$

where ε denotes the length of edge (u_2, v) in P_2. Thus P'_2 and R' separate at a node that we denote by u'_2, which has depth $\alpha + \beta_2 - \varepsilon$ in P'_2 (the same as u_2 in P_2). When $P'_1 \cup R'$ is a crack, because it is isomorphic to $P_1 \cup R$, the first node in the suffix common to P'_1 and R' must be at the same depths in P'_1 and R' as t in P_1 and R. We call this node t'. Because the depth of t in R is strictly larger than $\alpha + \beta_2 - \varepsilon$, the same holds for the depth of t' in R', implying that t' must be a strict descendant of u'_2.

Now recall that v' is the first node that P'_1 and P'_2 have in common after separating at u'. There are two possibilities regarding its position in P'_2 relative to u'_2. First consider the case where v' is an ancestor of u'_2 along P'_2 (including $v' = u'_2$). In this case v' is a strict descendant of u' in the prefix common to P'_2 and R', and at the same time v' is a node in P'_1. But this contradicts both possible relations between P'_1 and R': P'_1 and R' forming a wishbone and separating at u', and P'_1 and R' forming a crack by separating at u' and joining in t' (which, as we showed, must be a strict descendant of u'_2, implying $t' \neq v'$). The other case to consider is that of v' being a strict descendant of u'_2 along P'_2, while being a strict ancestor of v'_2 along P'_2 (claim **C3**). That is, v' is strictly between u'_2 and v'_2 in P'_2. But this is impossible, as together with the fact that P_2 and P'_2 are isomorphic, it would imply the existence of a node strictly between u_2 and v in P_2, which is excluded by construction. Since all these possibilities lead to a contradiction, we conclude that **C3** is also impossible.

Finally, let us deal with claim **C4**: suppose that $v' = v'_2$, and that v' is a strict ancestor of v'_1 along P'_1. (See Fig. S9 to follow the argument below.) Recall that, because P'_1 and P'_2 are isomorphic to P_1 and P_2, they have lengths $\alpha + \beta_1 + \gamma$ and $\alpha + \beta_2 + \gamma$, respectively. Also recall that P'_1 and P'_2 separate at u' (with depth α in P'_1 and P'_2) and that the respective depths of v'_1 and $v' = v'_2$ in P'_1 and P'_2 are $\alpha + \beta_1$ and $\alpha + \beta_2$. Now, let $\delta > 0$ denote the length of the weighted path in P'_1 from v' to v'_1, implying that the weighted path in P'_1 from u' to v' has length $\beta_1 - \delta > 0$ (strictly positive because by construction $u' \neq v'$). Now define in N' a new root-leaf path S' that coincides with P'_1 and P'_2 along all their common prefix down to u' (of length α), then coincides with P'_1 along the weighted path in P'_1 from u' to v' (of length $\beta_1 - \delta$), and finally coincides with P'_2 along all its suffix from v' onwards (of length γ). As a result, S' and P'_1 have a common prefix down to v' and therefore $\mu(P'_1, S') \geq \alpha + \beta_1 - \delta$.

Let us now focus on the consequences on N of these definitions. Let S denote the unique root-leaf path in N isomorphic to S' (Lemma 16) and let s be the node at depth $\alpha + \beta_1 - \delta$ in P_1 (whose existence

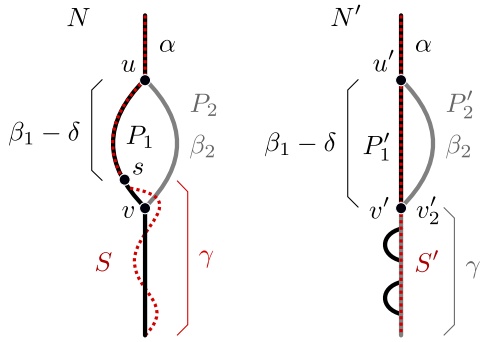


Figure S9. Illustration of the argument against claim **C4** in the proof of Lemma 18. P_1 and P_1' are in black and the portions of P_2 and P_2' not overlapping with P_1 and P_1' are in grey. Finally S and S' are represented by the dashed lines in red.

is guaranteed by the existence of node v' at the same depth in P_1'). Because the same holds for P_1', P_2' and S' in N', the following holds (Lemma 15):

$$\mu(P_1, P_2) = \mu(P_2, S) = \alpha < \alpha + \beta_1 - \delta \leq \mu(P_1, S),$$

meaning that the prefix of P_1 down to s is entirely in common with S (and therefore S and P_2 separate in u). But this implies that all the nodes that S has in common with P_2 after separating from it at u must be strict descendants of s, as otherwise P_1 and P_2 would have nodes in common between u and v, which is excluded by construction.

Now note that, because P_2' is isomorphic to P_2, and because the suffix of P_2 starting in v contains k edges, then the suffix of P_2' starting in $v' = v_2'$ (the node at the same depth in P_2' as v in P_2) also contains exactly k edges. But this suffix coincides with that of S' starting in $v' = v_2'$. Then, also the suffix of S starting in s (the node at the same depth in S as v' in S') contains exactly k edges. This observation, together with the fact that all the nodes that S has in common with P_2 after separating from it must be strict descendants of s, allows us to show that for every wishbone or crack W contained in $P_2 \cup S$ and containing S, there exists in N' a wishbone or crack W' isomorphic to W. This is trivial when W is a wishbone: because P_2 and S are root-leaf paths to the same taxon, the only wishbone W contained in $P_2 \cup S$ and containing S is S itself, for which $W' = S'$. This is also trivial when W is a closed crack, as we have already proved that N and N' must have the same closed cracks (up to isomorphism). It remains the case where W is an open crack. In this case, it can be written as $W = S \cup Q$, where Q is a root-leaf path contained in $P_2 \cup S$, separating openly from S at some internal node in S. Because all the nodes that S and P_2 have in common after separating at u must be strict descendants of s, also the first node in the longest suffix common to S and Q must be a strict descendant of s, meaning that this suffix must contain strictly less than k edges. Then, by inductive hypothesis, there must be a crack W' in N' that is isomorphic to $W = S \cup Q$. Because for every wishbone or crack W contained in $P_2 \cup S$ and containing S, there exists in N' a wishbone or crack W' isomorphic to W, then P_2' and S' must intersect each other at the same depths as P_2 and S (by Lemma 10). But this leads to a contradiction, as the node at depth $\alpha + \beta_1 - \delta$ in S', that is $v' = v_2'$, belongs to both S' and P_2', whereas the node at the same depth in S, that is s, only belongs to S.

We have thus proved that each of **C1**-**C4** leads to a contradiction. The fact that neither **C2** nor **C3** can hold implies that $v' = v_2'$. This, together with the fact that neither **C1** nor **C4** can hold, implies $v' = v_1'$. Thus, $v' = v_1' = v_2'$, which, as explained in the introduction of this proof, implies that $P_1' \cup P_2'$ is a crack isomorphic to $P_1 \cup P_2$, and thus the lemma follows. □

Proposition 2 follows from Corollary 4 and Lemma 18.

Networks with inheritance probabilities and likelihood-based reconstruction

As described in the main text, the ML framework we consider [2, 32, 33, 38] not only models edge lengths, but also inheritance probabilities. The latter provide, for each reticulate edge in a network N, the probability that a random tree T displayed by N includes that edge (i.e., that inheritance follows that edge). The inheritance probabilities determine, for each tree $T \in \mathcal{T}(N)$, an associated probability $\mathbf{Pr}(T|N)$.

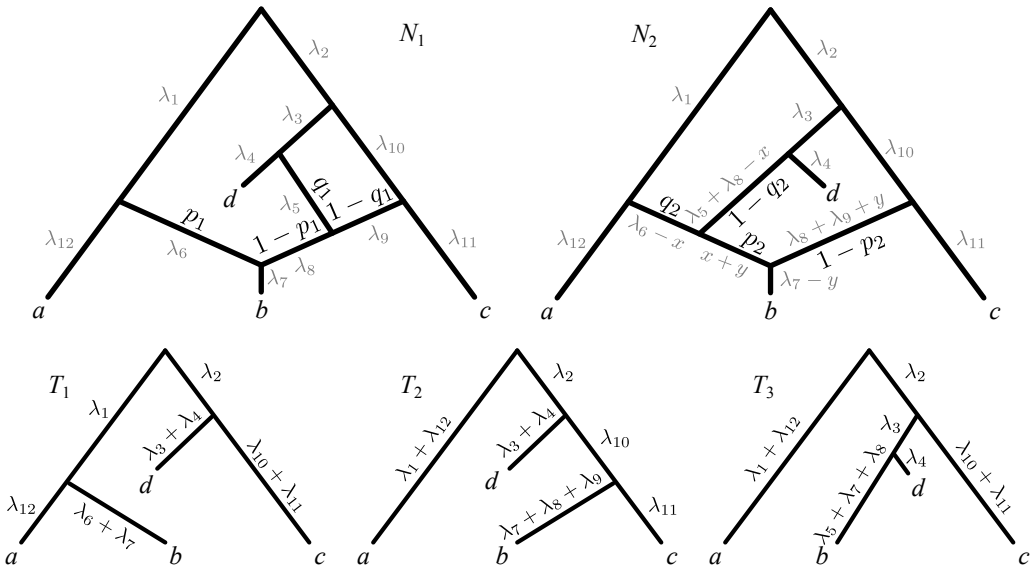
Unfortunately, including inheritance probabilities does not solve identifiability problems: in Fig. S10 we show an example of two phylogenetic networks N_1, N_2 with edge lengths and inheritance probabilities that cannot be distinguished on the basis of the trees they display and associated probabilities. By setting the edge lengths as shown, and the inheritance probabilities so that $p_2 = 1 - (1 - p_1)(1 - q_1)$ and $q_2 = p_1/p_2$, it is easy to check that $\mathbf{Pr}(T_1|N_1) = \mathbf{Pr}(T_1|N_2)$, $\mathbf{Pr}(T_2|N_1) = \mathbf{Pr}(T_2|N_2)$ and $\mathbf{Pr}(T_3|N_1) = \mathbf{Pr}(T_3|N_2)$. Thus, for every assignment of edge lengths and inheritance probabilities to N_1, there exist corresponding assignments to N_2 that make the resulting networks display the same trees, with the same edge lengths and the same probabilities of being observed. Because $\mathcal{T}(N_1) = \mathcal{T}(N_2)$, and $\mathbf{Pr}(T|N_1) = \mathbf{Pr}(T|N_2)$ for any T displayed by these two networks, it is easy to see that

$$\prod_{i=1}^{m} \sum_{T \in \mathcal{T}(N_k)} \mathbf{Pr}(A_i|T)\mathbf{Pr}(T|N_k).$$

is the same for $k = 1$ or $k = 2$, that is, the likelihoods of N_1 and N_2 are identical regardless of the data. In other words, no method based on this definition of likelihood will be able to discriminate between them.

The example above is not an exception: for any two distinct indistinguishable networks (i.e., with $\mathcal{T}(N_1) = \mathcal{T}(N_2)$), it is possible to provide assignments of inheritance probabilities to their reticulations, so that not only these networks display the same trees, but that also the probabilities associated to the trees they display are identical.

Moreover, we can extend the notion of indistinguishability, as well as that of canonical form, to networks with inheritance probabilities. Results entirely analogous to those we presented in this paper will then hold. We will not prove any of this here, as it lies beyond the scope of the present study.



$$\mathbf{Pr}(T_1|N_1) = p_1 \qquad \mathbf{Pr}(T_2|N_1) = (1-p_1)(1-q_1) \qquad \mathbf{Pr}(T_3|N_1) = (1-p_1)q_1$$

$$\mathbf{Pr}(T_1|N_2) = p_2q_2 \qquad \mathbf{Pr}(T_2|N_2) = 1-p_2 \qquad \mathbf{Pr}(T_3|N_2) = p_2(1-q_2)$$

Figure S10. An example showing that the ML framework considered in the main text is also subject to identifiability problems: the two networks N_1 and N_2 (*top*) with edge lengths (gray) and inheritance probabilities (black), display the same trees T_1, T_2, T_3 (*middle*), with the same associated probabilities (*bottom*), when $p_2 = 1-(1-p_1)(1-q_1)$ and $q_2 = p_1/p_2$.