An Alternative Construction of Internodons: The Emergence of a Multi-level Tree of Life

Samuel Allen Alexander*
Arie de Bruin†
D. J. Kornet‡

2013

Abstract

Internodons are a formalization of Hennig's concept of species. We present an alternative construction of internodons imposing a tree structure on the genealogical network. We prove that the segments (trivial unary trees) from this tree structure are precisely the internodons. We obtain the following spin-offs. First, the generated tree turns out to be an organismal tree of life. Second, this organismal tree is homeomorphic to the phylogenetic Hennigian species tree of life, implying the discovery of a multi-level tree of life: this phylogenetic tree can be obtained by zooming out from the organismal tree, or conversely, the organismal tree of life can be generated by expanding the phylogenetic nodes into unary trees. Finally, the definition of the organismal tree allows an efficient algorithmic transformation of a given genealogical network into its corresponding phylogenetic species tree of life. The latter will be presented in a separate paper.

1 Introduction

According to Hennig, the hierarchical classification of organisms into species, genera, families, orders, etc., is based on a phylogenetic tree of species satisfying Woodger's formal definition of a hierarchy [14].[1] In essence, Woodger's

*The Ohio State University. Email: alexander@math.ohio-state.edu
†TU Delft
‡Leiden University

[1] "If, therefore, the relationships between the elements of a hierarchy are represented by unidirectional arrows, then according to Woodger's definition: (1) The point of one, and only one, arrow can lie in each element of the hierarchy, whereas several arrows may arise from it. (2) There is one, and only one, element from which arrows emanate but to which no arrows lead. Woodger and Gregg call this element the "beginner". (3) All elements to which an arrow leads, and which therefore lie at an arrow tip, are connected with the beginner by an arrow or a sequence of arrows." [5]

hierarchies are rooted trees. As Hennig states: "The 'phylogenetic tree' is only a different, sketch like form of presenting the hierarchic system" [5].

Hennig builds the notion of the phylogenetic tree of life on a very specific conception of species: "Such a picture of phylogenetic relationships can be a system of hierarchic type only if in its plan of construction the species is regarded as the unit that undergoes division. This is possible only if two successive processes of species cleavage are assumed to be the temporal delimitation of the existence of a species" (ibid. p. 64).

A species cleavage was sketched as follows: "If we could determine the genealogical relationships among all individuals over a long period of time, and present these graphically, we would find gaps in the structure of the relationships. These gaps divide complexes of individuals, which we call species, from one another" (ibid. p. 18). Under reference to splitting events as nodes, Hennig's concept of species was named the 'internodal species concept' by Nixon and Wheeler [11].[2]

Hennig wished to anchor his methodology of tree construction formally, and he managed to do so for phylogenetic trees by basing them on Woodger's formal definition of hierarchies. But he described his concept of species only verbally and graphically, for instance, by indicating a splitting event with just a solid triangle without formalizing this (ibid. Figs. 4, 6).

Hennig's concept of species has been formalized by Kornet [7] via equivalence classes based on a conspecificity relation **INT** on the organisms of a genealogical network. This formalization unveiled an implied dependence, in Hennig's informal split-triangles, upon infinity (See also [1]). These equivalence classes were later called internodons[3] [9]. This formalization amounts to a transformation of a genealogical network into a phylogenetic tree of species.

In the present paper, we take this project a step further by constructing an organismal tree to be nested within the phylogenetic tree of Hennigian species. This organismal tree of life not only is homeomorphic to the phylogenetic tree of Hennigian species but also defines it rigorously meeting equally well Woodger's formal criteria for hierarchies. To give an impression of the transformation of a genealogical network to such an organismal tree structure, we first present a naive example, in the sense that the delimitation of the species is stipulated a priori.

So, suppose we have a genealogical network in which the species are given (Fig. 1a), from which the phylogenetic species tree (Fig. 1b) is derived. Taking for granted that no two organisms are born at precisely the same time (almost certain if we measure time in small enough units), species (save possibly for species that endure forever unsplit) have unique youngest and oldest specimens. More strongly, we can view a species as a discrete ordered list: the oldest specimen, the next oldest, and so on. In this sense, species are trivial organismal

[2]This practice fits into 'stem-based tree' terminology, not to be confused with 'node-based tree' terminology where the nodes represent the species. See [10] and [12] for discussion of mixing node- and stem-based terminology. Confusion about the nodal part of the name could be avoided by calling it the 'intersplittal species concept.'

[3]In hindsight perhaps better named "intersplittons".

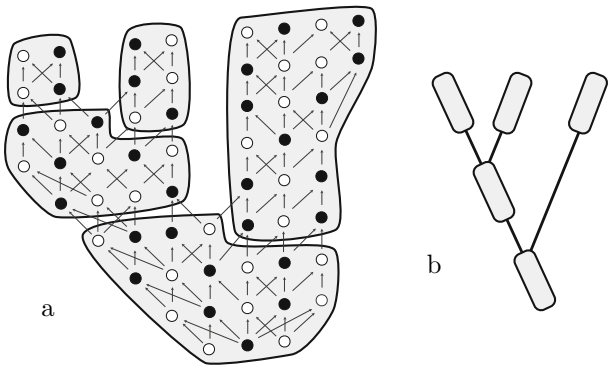

Figure 1: Binary phylogenetic species tree (open and closed circles denote male and female specimens). **a** A phylogenetic tree of five internodal species mapped on a genealogical network of organisms (a directed acyclic graph of their parental relation). **b** Node-based representation of the phylogenetic tree of five internodal species.

trees: unary trees, without branching. These trivial trees can be joined according to the species they represent: when one old species gives way to two new, the old trivial tree can be attached to the two new, with edges directed from the youngest specimen of the former to the oldest specimens of the latter (Fig. 2a). If we zoom in on such a species tree, we see a tree of organisms (Fig. 2b).

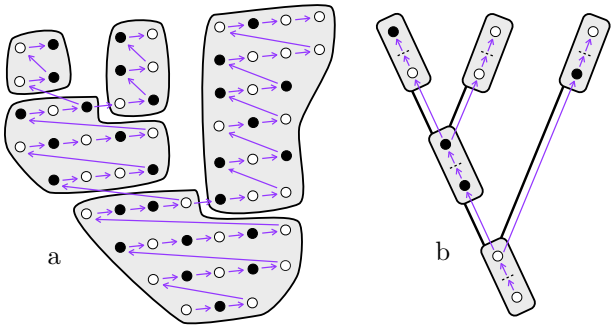

Figure 2: Unary/binary organismal tree nested in binary species tree. **a** Each species contains a unary tree of organisms ordered by birthdate. Last and first borns of consecutive species connect the unary trees into a unary/binary organismal tree. **b** Nodes show the first and last born of one internodal species. The unary/binary organismal tree is nested in the binary species tree.

As it assumes species already known (which marks the naivety of the example), the above reasoning, by itself, is useless for species delimitation in a genealogical network. But in this paper, we will construct these species from scratch. We will obtain a zoomable organismal/species tree structure much like

the example described above, assuming nothing more than tokogenetic information[4] of a given organismal network, the permanency of its splits and the total order induced by unique birthdates.

We will show that the 'trivial' unary organismal trees emerging in this construction are exactly Kornet's internodons. In this way, we arrive at a more perspicuous alternative definition of the internodon concept as well as a more insightful proof that internodons are indeed equivalence classes. A promising feature of this construction is that it is algorithmic in nature, which enables us to construct efficient software calculating trees of life out of genealogical networks via organismal trees.

A few words on the basic graph structure on which our paper will be built. Following [4], we view the biosphere as a directed graph G whose vertices are the individual organisms, with an edge directed from vertex x to vertex y if x is a parent of y. Following [1], we allow that G may be infinite. We assume organisms have unique birthdates, ancestors are born before their descendants, and that only finitely many organisms were born before any time t. The definition below formalizes this and introduces some terminology and notation used throughout the paper:

Definition 1. *1. A graph is called* birthdated *if every vertex has a unique real number associated to it, called its* birthdate[5]*. We write $x < y$ if x's birthdate is smaller (i.e., earlier) than y's, in which case we also say x is an* elder *of y, that x is* older *than y, or that y is* younger *than x.*

2. A birthdated graph is called downward finite *if it has, for each real number t, only finitely many vertices older than*[6] *t.*

3. Throughout this paper, the symbol G will denote a downward finite birthdated directed acyclic graph, where the vertices denote organisms, and the relation defined by the arrows is called the parent *relation, with the additional proviso that parents are always older than their children.*

4. For any members x, y of G, we say x is an ancestor *of y (equivalently, that y is a* descendant *of x) if $x \neq y$ and there is a sequence $x_0, \ldots, x_n$ of members of G such that each x_i is a parent of x_{i+1} (for $0 \leq i < n$),*

[4]That is, information pertaining to theparenthood relation between organisms (as defined by the genealogical network).

[5]It would be more appropriate, but less natural, to speak of birth moments rather than birthdates. We impose the constraint that no two members of the population have the same birthdate: a constraint that makes more sense the finer we divide time. Formally, the role birthdates play is to extend the partial order of ancestorhood into a total order, and for this purpose, we need not worry about the precise details of what exactly defines an organism's exact moment of birth.

[6]If, in clause 1 of this definition, we would have stipulated birthdates to be natural numbers instead of reals, then downward finiteness would follow immediately from uniqueness of birthdates. In the sequel of this paper, it will transpire that birthdates are needed only insofar as they impose a topological order on the graph, and natural numbers are sufficient for this purpose. However, we chose to stick to defining birthdates as real numbers because this is closer to the day-to-day meaning of this notion.

$x_0 = x$, *and* $x_n = y$ *(in short: if there is a nonempty directed path from* x *to* y*).*

The paper has the following structure. The construction of the organismal species tree hinges on a transformation of the parental relation defining the genealogical network (cf. Fig. 1a) into a derived relation to be called the undirparent relation which gives rise to organismal tree structures like the ones in Fig. 2a, b. This transformation is a special case of the so-called parenthood reconstruction method to be discussed in Section 2 and will be worked out for our special case in Section 3. In this section, we will also show that this indeed results in a natural forest or tree structure on the biosphere, called the undirtree. In Kornet's internodon construction, those organisms that have a descendant that has more than one parent, called **SD** organisms, play a vital role. Section 4 will be devoted to a study of such organisms.

By then, we have acquired sufficient material to give an alternative definition of Kornet's internodons, looking like the unary trees comprising the pre-given species in Fig. 2. This will be the subject of Section 5. Section 6 is devoted to a proof that these two internodon definitions are equivalent. In Section 7, we revisit the **SD** property and use infinitary combinatorial methods to show that in a formal sense, non-**SD** organisms are negligible. The final section is devoted to conclusions and future research.

2 The Parenthood Reconstruction Problem: Can Parenthood be Reconstrued from Ancestorhood?

The approach followed in this section is inspired by the observation that the basis of Kornet's internodon construction is the notion of path-connectedness which can naturally be interpreted as an ancestry-like relation. In a rather elaborate way, Kornet derived from this relation a disjoint cover consisting of internodons, thus arriving at a species tree. In seeking a shortcut, the route taken in the current paper is to derive from that ancestry-like relation a parent-like relation, arriving at an organismal tree of life. Since deriving parenthood relations from ancestral relations is far from simple, in this section we treat this general problem, the *parenthood reconstruction problem*, first.

We assume full ancestral knowledge of a population, meaning that we know the members of the population and given any two members, and we know whether or not one is an ancestor of the other (as in Definition 1). We further assume knowledge of birthdates.

The question we now wish to answer is which ancestor relations are parental relations as well. The following observation shows that without additional information, perfect parenthood reconstruction is hopeless.

Lemma 1. *There are two populations with different parental relations and the same ancestral relations. Thus, given just a population's ancestral relations, we*

cannot perfectly reconstruct its parental relations.

Proof. (See Fig. 3) Let a, b, c be three hypothetical organisms. Let X be the population with members a, b, c where a is the lone parent of b, and b is the lone parent of c. Let Y be the same population, with one difference: a and b co-parent c. It is easy to check X and Y have identical ancestral relations. □

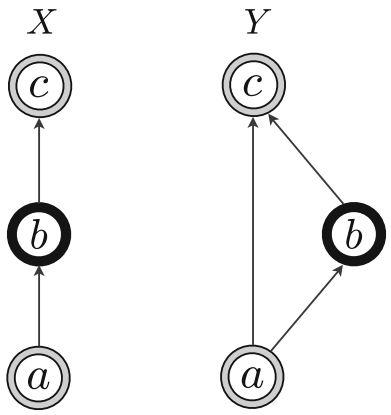

Figure 3: Proving Lemma 1: a being a grandparent of c allows multiple patterns of parenthood between a, b, and c.

We will give a parsimony-based imperfect solution to the parenthood reconstruction problem assuming ancestral relations that make some sense to begin with. For the remainder of this section, let X be a birthdated population.

Definition 2. *A binary relation $\prec_A$ on X is a* plausible ancestorhood *if the following conditions hold.*

1. *Ancestors are older than descendants: whenever $x \prec_A y$, x's birthdate is smaller than y's.*
2. *Transitivity: if $x \prec_A y$ and $y \prec_A z$, then $x \prec_A z$.*
3. *No organism has infinitely many ancestors: for any x, there are only finitely many y such that $y \prec_A x$.*

Definition 3. *Let $\prec_A$ be a plausible ancestorhood on X. By the* parsimoniously reconstructed parenthood*, we mean the binary relation $\prec_P$ such that for every $x, y \in X$, $x \prec_P y$ if and only if the following hold.*

1. *$x \prec_A y$.*
2. *There is no z such that $x \prec_A z$ and $z \prec_A y$.*

Proposition 2. *Let $\prec_A, \prec_P$ be as in Definition 3.*

1. *$\prec_P$ is consistent with $\prec_A$, by which we mean that $\forall x, y \in X$, $x \prec_A y$ if and only if there is a sequence $x = x_0, \ldots, x_n = y$ such that $x_i \prec_P x_{i+1}$ for each $i = 0, \ldots, n-1$.*

2. *$\prec_P$ is as small as possible: if $\prec_{P_0}$ is any binary relation consistent with $\prec_A$ in the above sense, then $\prec_P \subseteq \prec_{P_0}$ (i.e., $\prec_P$ is a subrelation of $\prec_{P_0}$, or equivalently, whenever $x \prec_P y$, then it follows that $x \prec_{P_0} y$).*

Because of the generality of Proposition 2, we can apply it freely to any ancestorhood notion we like, as long as that notion is formally plausible, as in Section 3 where a derivative ancestorhood notion will be obtained by varying the properties of actual ancestorhood.

Proof of Proposition 2. We divide the proof into two claims.

(1) First, we prove $\prec_P$ is consistent with $\prec_A$, in the sense defined above.

(See Fig. 4a) First suppose $x \prec_A y$, we will show there is a sequence $x = x_0, \ldots, x_n = y$, each $x_i \prec_P x_{i+1}$. If $x \prec_P y$, this is trivial (let $x = x_0$, $x_1 = y$). If not, there must be some z such that $x \prec_A z$ and $z \prec_A y$. If $x \prec_P z$ and $z \prec_P y$, we are done (let $x = x_0$, $z = x_1$, $y = x_2$). If not, say $x \not\prec_P z$ (the other case is similar), then there must be some z' with $x \prec_A z'$ and $z' \prec_A z$. This process continues... it cannot continue forever, lest y have infinitely many $\prec_A$ ancestors (distinct since ancestors are older than descendants), violating the plausibility of $\prec_A$.

Conversely, assume there is a sequence $x = x_0, \ldots, x_n = y$, each $x_i \prec_P x_{i+1}$, we will show $x \prec_A y$. Part of the definition of $x_i \prec_P x_{i+1}$ is that $x_i \prec_A x_{i+1}$. So each $x_i \prec_A x_{i+1}$, and by transitivity, $x \prec_A y$.

(2) Second, we prove $\prec_P$ is as small as possible.

Suppose $\prec_{P_0}$ is consistent with $\prec_A$, we will show $\prec_P \subseteq \prec_{P_0}$. Assume, for sake of contradiction, there are $x, y \in X$ such that $x \prec_P y$ and $x \not\prec_{P_0} y$. Since $x \prec_P y$, in particular, $x \prec_A y$. Thus, since $\prec_{P_0}$ is consistent with $\prec_A$, there is a sequence $x = x_0, \ldots, x_n = y$ with each $x_i \prec_{P_0} x_{i+1}$ (see Fig. 4b). We must have $x_1 \neq y$, by the assumption that $x \not\prec_{P_0} y$. Because $\prec_{P_0}$ is consistent with $\prec_A$, the subsequence $x_1 \prec_{P_0} \cdots \prec_{P_0} x_n = y$ forces $x_1 \prec_A y$. Letting $z = x_1$, this contradicts $x \prec_P y$ (Definition 3 part 2). □

3 Introduction to the Wondrous Undirworld: from Networks to Trees

Now, we will apply the construction from the previous section to the path-connected relation as given in Kornet's earlier work.

Recall (Definition 1) that x is an ancestor of y precisely if G has a nonempty directed path from x to y. We can strengthen this characterization slightly. The following lemma is trivial, but it plays an important role in the definition of internodons.

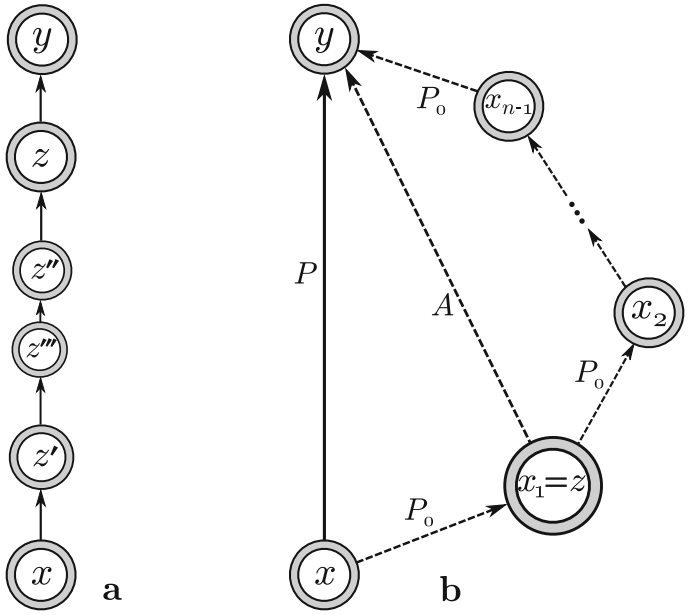

Figure 4: Proving optimality of the parsimoniously reconstructed parenthood.

Lemma 3. *(Compare underlined words with Definition 4 below) If $x \neq y$ are organisms, x is an ancestor of y precisely if G has a nonempty directed path from x to y that avoids x's elders.*

Proof. Follows from the assumption that parents are born before their children. □

The virtue of Lemma 3 is that it lends itself well to variation. From it, we obtain a definition by variation of what Kornet et al called [9] the *path-connected* relation, called **CNB** in the Appendix of [7].

Definition 4. *(Compare underlined phrases with Lemma 3 above) If $x \neq y$ are organisms, x is an undirected ancestor of y precisely if G has a nonempty undirected path from x to y that avoids x's elders.*

Figure 5a depicts a typical (directed) path as in Lemma 3, versus (Fig. 5b) a typical (undirected) path as in Definition 4. We have underlined words in Lemma 3 and Definition 4 to emphasize their differences.

One advantage of such definitions by variation is that they permit adjacent definitions by analogy. Thus, y is an *undirected descendant* of x if x is an undirected ancestor of y. For brevity, we use abbreviations like *undirancestor* and *undirdescendant* (the *undir* prefix can be pronounced "under").

Lemma 4. *Undirancestry is plausible in the sense of Definition 2.*

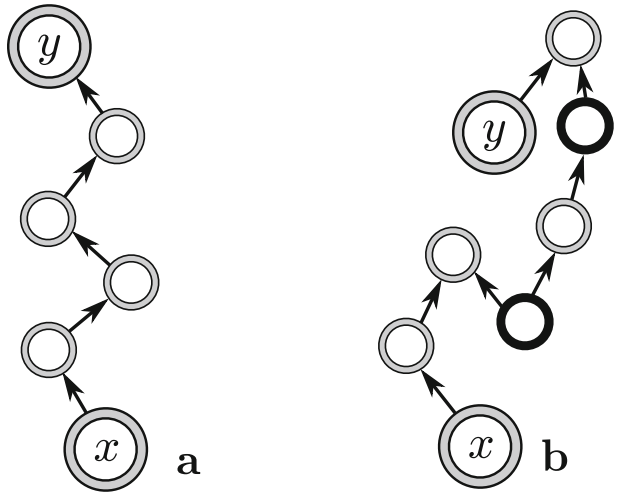

Figure 5: Directed and undirected paths avoiding x's elders.

Proof. We divide the proof into three parts corresponding to the three parts of Definition 2.

1. Undirancestors are older than their undirdescendants, because by definition, the defining path from an undirancestor a to each of its undirdescendants d avoids a's elders so that, more in particular, each d is not an elder of a. Because every organism in G has a unique birthdate, each d must be younger than a.

2. (Transitivity) If a is an undirancestor of b and b is an undirancestor of c, then there is an undirected path π from a to b avoiding a's elders, as well as an undirected path π' from b to c avoiding b's elders. It follows from (1) that π' also avoids a's elders. Thus, the concatenation of π and π' is an undirected path from a to c avoiding a's elders.

3. Suppose organism a has infinitely many undirancestors. Then, by (1), there are infinitely many organisms older than a. But in that case, there are infinitely many organisms born before a's birthdate, which contradicts the fact that G is downward finite (Definition 1 part 2).

□

The above Lemma justifies the Definition below, which is based on Proposition 2.

Definition 5. *Following Definition 3, x is an* undirparent *of y if the following conditions hold.*

1. x is an undirancestor of y.

2. *There is no organism z such that x is an undirancestor of z and z is an undirancestor of y.*

Figure 6a shows parenthood relations, and Figure 6b shows undirparenthood relations in a small population. By Proposition 2, undirparenthood is consistent with undirancestorhood. For motivation, recall the introduction. Under Hennig, individual species can be viewed (as we do in Figure 2a) as unary trees joined where speciation occurs to form a compound tree of life. The following theorem shows that undirparenthood formalizes this. It will turn out that the undirtree construction gives rise to a segmentation of this tree that forms the basis of the formal internodal species concept, which is well motivated [7] [9] [8]. That the same notion arises in such different ways increases its robustness, in the mathematician's view.

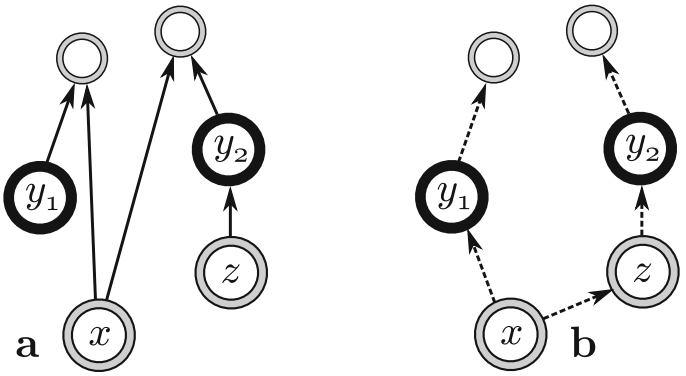

Figure 6: Parental relations (**a**) and undirparental relations (**b**).

Theorem 5. *Let G' be the digraph (directed graph) whose vertex set is the set of organisms, with an edge directed from x to y precisely when x is an undirparent of y. Then G' is a forest. If the biosphere is monophyletic (i.e., if all life descends from a common ancestor), G' is a tree.*

Proof. (See Fig.7) It suffices to show every organism has at most one undirparent. This will hinge on our assumption that no two organisms share a birthdate.

Let y be an organism and, for sake of contradiction, assume y has distinct undirparents x_1 and x_2. Relabeling if necessary, say x_1 is born before x_2. Because x_1 and x_2 are undirparents of y, in particular, they are undirancestors of y, so there are nonempty undirected paths π (from x_1 to y) and ρ (from x_2 to y) such that π avoids x_1's elders and ρ avoids x_2's elders. Then, $\pi \frown \rho^{reverse}$ is a nonempty undirected path from x_1 to x_2, avoiding x_1's elders[7]. This shows x_1 is an undirancestor of x_2. Letting $z = x_2$, this violates Definition 5 Part 2.

If the biosphere is monophyletic, then G' is connected, hence a tree. □

[7]Here, $\frown$ denotes concatenation and $\rho^{reverse}$ stands for the path ρ travelled in the opposite direction. More formally, if $\rho = a_1, \ldots, a_n$ denotes an undirected path (where the a_i are nodes), then $\rho^{reverse}$ is the path $a_n, ..., a_1$.

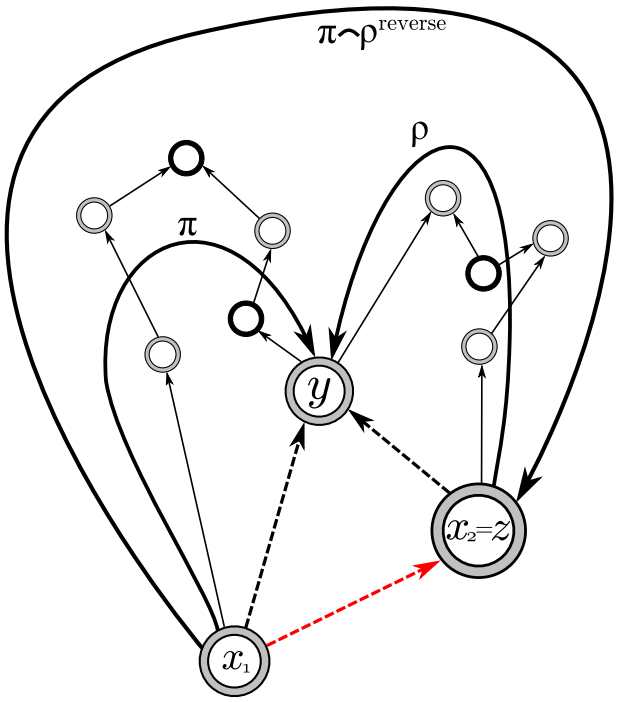

Figure 7: Proving that undirparenthood is a forest.

Theorem 5 provides a foothold for doing with DAGs what is normally done with trees (for example, defining monophyletic groups, closer to cladograms), but that is not the primary focus of this paper.

Corollary 6. *Any time y has an undirparent x, x is the youngest undirancestor of y.*

Call the graph G' of Theorem 5 the *undirforest*, or the *undirtree* if the biosphere is monophyletic. Figure 8a shows an initial segment of a population, and Figure 8b shows the corresponding undirtree. We have arrived at a tentative candidate for (an organism-level version of) the tree of life. On our own, we have little justification to call this tree such, but it will turn out that it is identical to a zoomed-in version of the tree of internodons. Before we prove as much, we must first analyze the structure of G'.

We adopt the following notation from [9].

Definition 6. *For any organisms x, y, write $x(\boldsymbol{PC}_{\geq x})y$ if there is a (possibly empty) undirected path from x to y avoiding x's elders. Thus, $x(\boldsymbol{PC}_{\geq x})y$ if either $x = y$ or x is an undirancestor of y. Write $x \leqq y$ if $x = y$ or x is born before y.*

Lemma 7. *For any organisms $x \leqq y$, $x(\boldsymbol{PC}_{\geq x})y$ if and only if there is a directed path from x to y in the undirforest G'.*

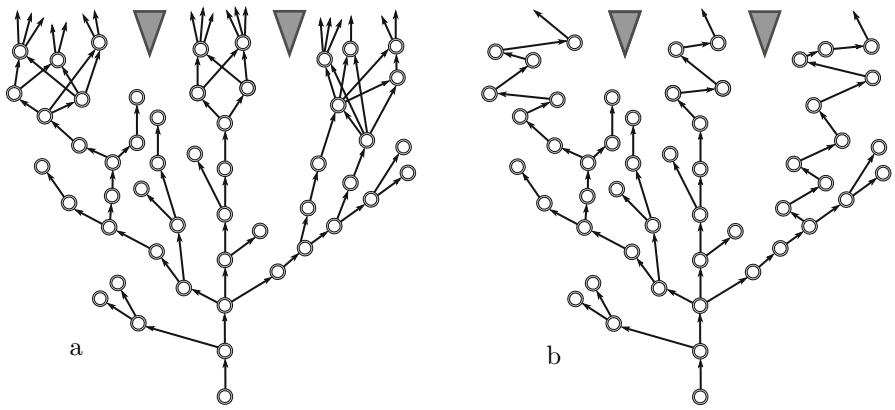

Figure 8: **a** A population (a network). **b** The corresponding undirpopulation (a tree). Triangles in the graph denote permanent splits.

Proof. Immediate from the fact (Proposition 2) that undirparenthood is consistent with undirancestry. □

4 The SD Property: Having a Sexual Descendant

A problem with the Hennigian notion of cleavages as speciation events is that it leads to unrealistically small species, for instance, via organisms that do not bear offspring. In the most extreme case, a single childless organism technically induces a permanent split[8]. One such infertile specimen would suffice to terminate its parents' species and generate two new species (one consisting solely of that single organism), if Hennig's notion is not amended to take this into account.

A general method for reducing the above problem is as follows (∗): formally define some notion of outsider organism, in such a way that the property of being an outsider is inherited by descendants (descendants of an outsider should be outsiders). Having done this, one may temporarily discard all outsider organisms; partition the remaining non-outsiders into species; and finally, place outsider organisms in the species of their most recent non-outsider ancestor. The only question, then, is how to define outsider organisms. There are many possibilities. Effectively, Kornet's [7] outsiders, called non-**SD** organisms (see Definition 7), are the organisms that have no descendant with more than one parent. Thus, to be non-outsider, one's descendants (and those descendants' parents) must form a proper genealogical subnetwork, not simply a tree. For a general notion of outsider, the construction (∗) might significantly distort the network (and there might be a significant amount of outsider organisms with no

[8]See [13] for more discussion on the problem of short-lived internodons.

non-outsider ancestors). But if we use Kornet's **SD** notion, we can give a very rigorous mathematical argument that both the distortion, and the number of non-**SD** organisms without **SD** ancestors, are negligible. (We do precisely that in Section 7.)

Definition 7. *An organism x in G has the* sexual descendant property *(or the* **SD** *property, or more simply x* is **SD***) if x has a descendant y such that y has at least two parents in G. If x is not* **SD***, an organism y is an x-*test organism *if the following hold.*

1. *y is* **SD***.*
2. *y is an ancestor of x.*
3. *No y' younger than y satisfies (1) and (2).*

Figure 9 shows a partially cleistogamous population. Non-**SD** members, indicated by smaller vertices, form non-**SD** clusters, indicated by dashed lines.

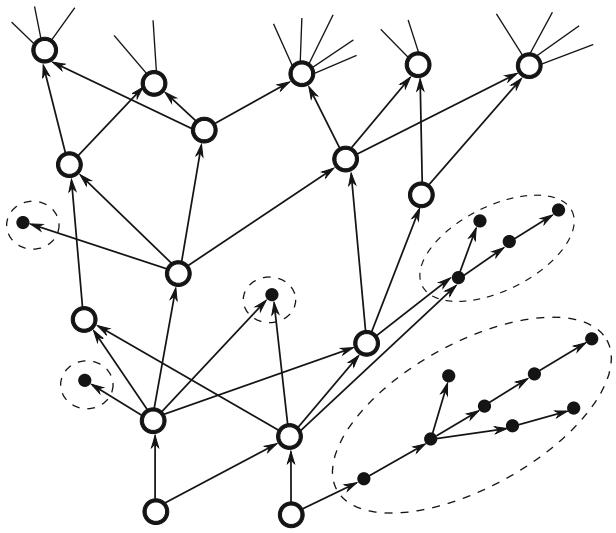

Figure 9: A partially cleistogamous population, with non-**SD** clusters indicated by *dashed* lines.

The internodon construction to be discussed in Section 5 will be based on the **SD** subgraphs of the networks G and G'. One can see from Figure 9 that omitting non-**SD** nodes there leaves the structure of the **SD** part intact in the sense that if two **SD** nodes are connected by a directed path in G, this will also be the case in the **SD**-only subgraph of G. (This property follows from the observation that an ancestor of an **SD** node is itself **SD**.) On the other hand, it is not immediately clear that the same property holds for the undirgraph G'. To prove that this is indeed the case (Proposition 13), we need a few preparatory technical lemmas.

Lemma 8. *Suppose x is a non-***SD** *organism. There is at most one x-test organism.*

Proof. Immediate from the assumption that no organisms share a birthdate and part 3 of Definition 7. □

Lemma 9. *An organism cannot have multiple non-***SD** *parents.*

Proof. Immediate from the definition of **SD**. □

Corollary 10. *The subgraph $G_0 \subseteq G$ of non-***SD** *organisms is a forest.*

Proof. Immediate from Lemma 9. □

Lemma 11. *Let x be non-***SD***. For any organism y, the following are equivalent.*

1. *x is an undirancestor of y.*
2. *x is an ancestor of y.*
3. *There is a unique nonempty directed path from x to y in G.*

Proof. ($1 \Rightarrow 2$) Suppose x is an undirancestor of y. This means there is a nonempty undirected path $x = p_0, \ldots, p_n = y$ from x to y avoiding x's elders (see Figure 10). Simplifying if necessary, we may assume this path does not repeat vertices. We claim each p_i is a parent of p_{i+1} (making x an ancestor of y). Assume not, then there is some i (we can take the minimum such i) such that p_i is a child[9] of p_{i+1}. Since the path avoids x's elders, $i > 0$. Thus, p_i has two parents p_{i-1} and p_{i+1}, distinct since the path does not repeat vertices. Because i was chosen minimal, p_i is a descendant of $p_0 = x$. This shows x has a descendant with multiple parents, violating the assumption that x is non-**SD**.

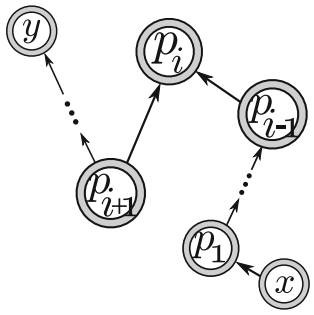

Figure 10: Proving that undirancestors that are non-**SD** are ancestors in the ordinary sense.

($2 \Rightarrow 3$) Suppose x is an ancestor of y. Then, there is a directed path π in G from x to y. In fact, all the vertices in π are non-**SD** (they are descendants

[9]Thus, for any two **SD** vertices x, y, if x is an undirparent of y, then there is an arc from x to y in $G'_{\mathbf{SD}}$.

of x, and if they were **SD**, that would make x itself **SD**), so lie in the non-**SD** forest of Corollary 10. It is a general fact from graph theory that there cannot be multiple directed paths connecting two vertices in a forest.

($3 \Rightarrow 1$) If there is a unique nonempty directed path π from x to y in G, π must avoid x's elders (since ancestors are born before descendants), and so since π is also an undirected path, π witnesses that x is an undirancestor of y. □

It can be shown from Lemma 11 that among non-**SD** organisms, parenthood and undirparenthood are the same relation.

Corollary 12. *1. Every undirdescendant of a non-***SD** *organism is non-***SD**.

2. Every undirancestor of an **SD** *organism is* **SD**.

Proof. (2) is the contrapositive of (1). To prove (1), let x be non-**SD** and let y be an undirdescendant of x. By Lemma 11, y is a descendant of x. If y were **SD**, then y would have a descendant z with multiple parents. This is impossible because z would also be a descendant of x. □

Proposition 13. *Let $G'_{\mathbf{SD}} \subseteq G'$ be the vertex-induced subdigraph of G' containing the* **SD** *organisms (i.e., $G'_{\mathbf{SD}}$ is the smallest subgraph of G' containing the* **SD** *organisms and containing every edge whose two endpoints are* **SD**[10]*).*

1. *$G'_{\mathbf{SD}}$ is a forest.*
2. *If $x, y \in G'_{\mathbf{SD}}$, then $x(\boldsymbol{PC}_{\geq x})y$ if and only if y is in the subtree of $G'_{\mathbf{SD}}$ rooted at x.*

Proof. (1) follows from Theorem 5, the rest of the proof is devoted to (2).

($\Rightarrow$) Assume $x(\mathbf{PC}_{\geq x})y$. By Lemma 7, there is a directed path π from x to y in G'. Since every nonterminal vertex in π is an undirancestor of y and y is **SD**, by Corollary 12, this path is actually a path in $G'_{\mathbf{SD}}$, placing y in the subtree of $G'_{\mathbf{SD}}$ rooted at x.

($\Leftarrow$) Immediate by Lemma 7. □

To illustrate Proposition 13, Figure 11a shows an undirtree G' (from Figure 8b) and Figure 11b shows the **SD**-part $G'_{\mathbf{SD}}$.

5 Internodons in the Undirworld[11]

We follow Kornet's idea of first partitioning the **SD** organisms into species, and then letting non-**SD** organisms fall into the species of their test organisms (Definition 7).

To avoid confusion, in the sequel, we will refer to the notion of internodon as defined in [9] as internodon$_K$, and to this notion as defined below as internodon$_A$ (for its motivation from [1]). After Section 6 where we will show that these definitions are equivalent, this distinction will not be needed anymore.

[10] Thus, for any two **SD** vertices x, y, if x is an undirparent of y, then there is an arc from x to y in $G'_{\mathbf{SD}}$.

[11] Note that the undirworld assumes unequal birthdates. See also Remark 21.

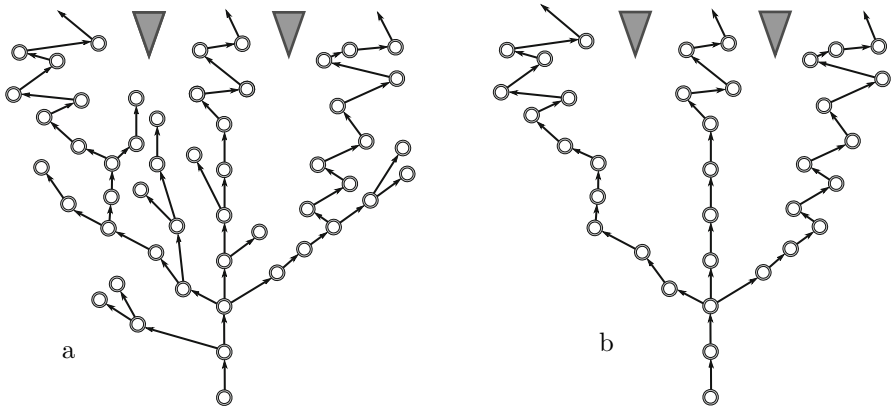

Figure 11: From a full undirtree to an **SD**-only undirtree. **a** An undirtree G' derived from the population in Figure 8a. **b** The corresponding **SD**-only part $G'_{\mathbf{SD}}$ of Proposition 12. *Triangles* denote permanent splits in the network.

Definition 8. *By a* segment *of a forest, we mean a maximal non-branching directed path, that is, a directed path whose vertices (except possibly a final vertex) have outdegree* 1 *and which is as long as possible with that property. By a* pre-internodon$_A$ *(see Figure 12), we mean a segment of the forest $G'_{\mathbf{SD}}$ of Proposition 13*[12]*.*

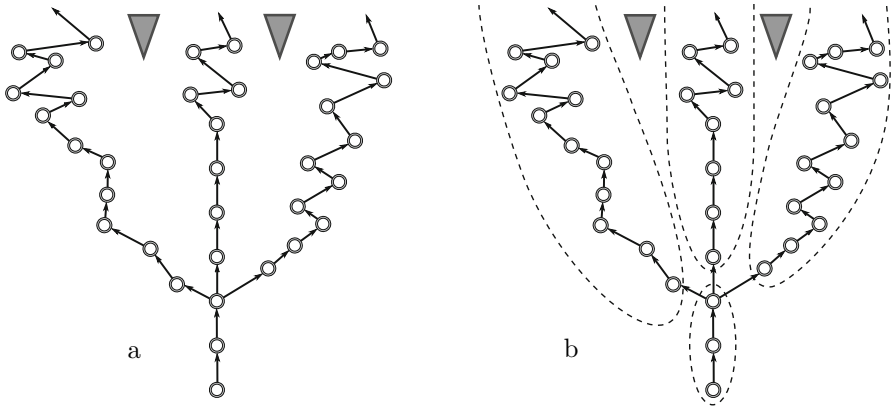

Figure 12: Pre-internodons$_A$. **a** The **SD** part of the undirpopulation from Figure 11. **b** The corresponding pre-internodons$_A$. *Triangles* denote permanent splits in the network.

Applying species terminology to pre-internodons$_A$, speciation occurs pre-

[12]Note that a non-final vertex in a segment of $G'_{\mathbf{SD}}$ may have outdegree > 1 in G', but must have exactly one **SD** undirchild.

cisely when an **SD** organism has multiple **SD** undirchildren. If an **SD** organism x has $n > 1$ **SD** undirchildren $y_1, \ldots, y_n$, then x's pre-internodon$_A$ goes extinct and n new pre-internodons$_A$ speciate, one for each of $y_1, \ldots, y_n$.

Lemma 14. *The pre-internodons$_A$ form a disjoint cover of the set of all* **SD** *organisms.*

Proof. A special case of the general fact that the segments of any forest form a disjoint cover of the vertices of that forest. □

We finally state the definition of internodon$_A$. In Section 6, we will recall the definition of internodon$_K$ from Kornet et al and prove that the two are equivalent.

Definition 9. *(See Figure 13) An* internodon$_A$ *is a set of the form $P \cup Q$, where P is a pre-internodon$_A$ and Q is the set of non-***SD** *organisms with test organisms in P.*

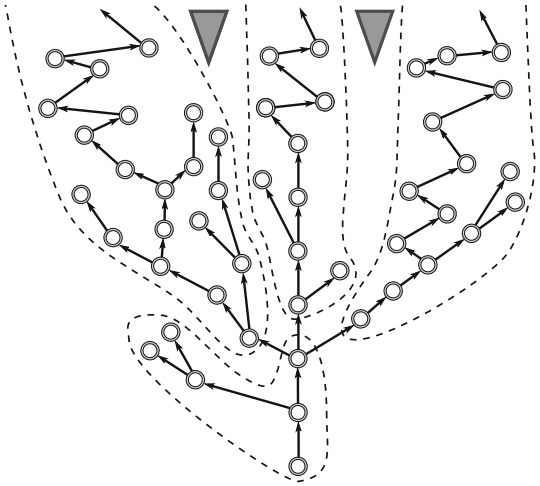

Figure 13: Internodons$_A$ in the population from Figure 12.

Lemma 15. *The internodons$_A$ form a disjoint cover of the set of all* **SD** *organisms combined with the set of all non-***SD** *organisms with test organisms.*

Proof. Immediate by Lemmas 8 and 14. □

6 Definitional Equivalence

In this section, we recall some definitions from Kornet et al [9]. We recall them in a simplified form, due to our assumption of distinct birthdates. The 1995 paper allows shared birthdates, which complicates definitions. In Remark 21, we revisit this issue.

Definition 10. *(Kornet et al, 1995) Assume x, y are organisms.*

1. *By $\mathbb{DYN}(x)$ (the* dynasty *of x), we mean $\{y : y$ is* **SD** *and $x(\boldsymbol{PC}_{\geq x})y\}$ (in the language of this paper, $\mathbb{DYN}(x)$ consists of x's* **SD** *undirdescendants along with x itself, if x is* **SD***).*

2. *By* **INTSD***, we mean the binary relation on the set of* **SD** *organisms defined so that for all* **SD** *organisms x and y,*

   (a) *If $x = y$, then x**INTSD**y.*

   (b) *If x is born before y, then x**INTSD**y holds if and only if:*

      i. *$y \in \mathbb{DYN}(x)$, and*

      ii. *For all $r \in \mathbb{DYN}(x)$ such that $y \geqq r$, $\mathbb{DYN}(r) = \{z \in \mathbb{DYN}(x) : z \geqq r\}$.*

   (c) *If x is born after y, then x**INTSD**y if and only if y**INTSD**x.*

Proposition 16. *Let I be a pre-internodon$_A$ and let $x \in I$. For any*[13] *organism y, $y \in I$ if and only if x**INTSD**y.*

Proof. Write $I = \{x_1, x_2, \ldots\}$ (possibly finite, possibly infinite) where each x_i is an undirparent of x_{i+1}.

($\Rightarrow$, see Figure 14) Assume $y \in I$, we will show x**INTSD**y. Assume x is born before y, the other case is similar. Write $x = x_j$, $y = x_k$, so $k > j$. The path $x_j, \ldots, x_k$ witnesses $y \in \mathbb{DYN}(x)$. Now suppose $r \in \mathbb{DYN}(x)$ and $y \geqq r$. By Proposition 13, r lies on the **SD** subtree rooted at x, so must be x_ℓ for some $k \geq \ell \geq j$; we claim $\mathbb{DYN}(r) = \{z \in \mathbb{DYN}(x) : z \geqq r\}$. Clearly, $\mathbb{DYN}(r) \subseteq \{z \in \mathbb{DYN}(x) : z \geqq r\}$. For the reverse inclusion, suppose $z \in \mathbb{DYN}(x)$ and $z \geqq r$. By Proposition 13, z is in the **SD** subtree rooted at x, so there is a directed undirparental path π from x to z. And π must pass through r, lest one of $x_j, \ldots, x_{\ell-1}$ have multiple undirchildren (contrary to Definition 8). A suitable restriction of π witnesses $z \in \mathbb{DYN}(r)$.

($\Leftarrow$, see Figure 15) Assume x**INTSD**y, we will show $y \in I$.

Assume y is younger than x, the other case is similar. Since $y \in \mathbb{DYN}(x)$, by Proposition 13, there is a directed undirparental path π from x to y. We claim there are no splitting points (organisms with multiple **SD** undirchildren) on π except possibly for y itself; this claim will prove $y \in I$ by maximality of I. Assume not: assume π passes through u, which has multiple **SD** undirchildren, before reaching y. Let r be the oldest **SD** undirchild of u and let r' be a different **SD** undirchild of u. Note $r \leqq y$ since π contains an **SD** child of u. Now, $\{z \in \mathbb{DYN}(x) : z \geqq r\}$ contains r' but $\mathbb{DYN}(r)$ does not contain r' (by Theorem 5 and Proposition 13). This violates the definition of x**INTSD**y. □

Corollary 17. INTSD *is an equivalence relation, and its equivalence classes are the pre-internodons$_A$.*

[13] Note that $y \in I$ implies that y is **SD**, which implication is also true if x**INTSD**y. Thus, the Proposition would not change if we wrote "for any **SD** organism y."

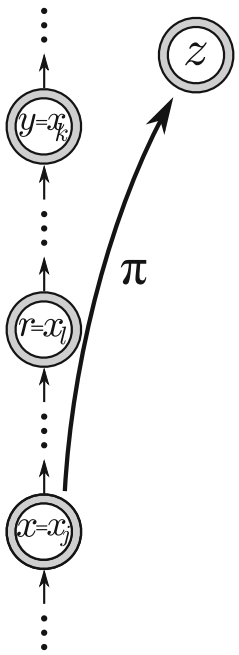

Figure 14: Proving that when two organisms x and y share a pre-internodon$_A$, they are necessarily **INTSD** related.

Proof. By Proposition 16, for any **SD** organism x, $\{y : x\mathbf{INTSD}y\} = \{y : y\mathbf{INTSD}x\}$ equals the pre-internodon$_A$ containing x. By Lemma 14, these pre-internodons$_A$ are a disjoint cover of the set of all **SD** organisms. This tells us **INTSD** is an equivalence relation and that the pre-internodons$_A$ are exactly the **INTSD** equivalence classes. □

On the basis of the first part of Corollary 17, the **INTSD** classes can aptly be called pre-internodons$_K$ (we would have made this a definition sooner, but it requires knowledge that **INTSD** is an equivalence relation); the second part of Corollary 17 then becomes a statement about equivalence of definitions: the pre-internodons$_K$ and the pre-internodons$_A$ are identical.

Definition 11. *(Kornet et al, 1995)*

1. *By* **WN**, *we mean the set of organisms that either are* **SD** *or have a test organism.*

2. *By* **INT**, *we mean the binary relation on* **WN** *defined so that for all organisms* $x, y \in \mathbf{WN}$, $x\mathbf{INT}y$ *if and only if* $x'\mathbf{INTSD}y'$, *where*

   (a) *If* x *is* **SD**, *then* $x' = x$, *otherwise* x' *is* x*'s test organism.*

   (b) *If* y *is* **SD**, *then* $y' = y$, *otherwise* y' *is* y*'s test organism.*

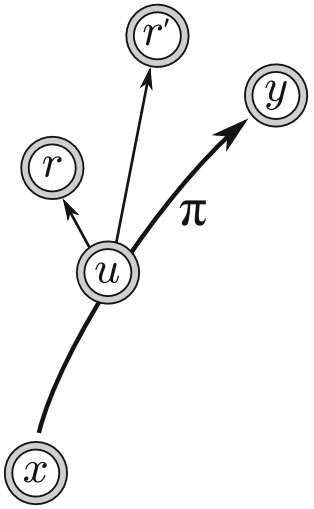

Figure 15: Proving that when two organisms x and y are **INTSD** related, they necessarily share a pre-internodon$_A$.

Proposition 18. *Let I be an internodon$_A$ and $x \in I$. For any organism y, $y \in I$ if and only if x**INT**y.*

Proof. Let x', y' be as in Definition 11 clause 2 and write $I = P \cup Q$ as in Definition 9 (so P is a pre-internodon$_A$). Since $x \in I$, it follows $x' \in P$. The following are equivalent:

$$\begin{aligned} y &\in I \\ y' &\in P && \text{(Definition 9)} \\ x'\mathbf{INTSD}y' &\text{(Proposition 16)} \\ x\mathbf{INT}y. &\text{(Definition 11)} \end{aligned}$$

□

Corollary 19. INT *is an equivalence relation and the equivalence classes are precisely the internodons$_A$.*

Definition 12. *(Kornet et al, 1995) By an* internodon$_K$*, we mean an* **INT** *equivalence class.*

Theorem 20. *(Definitional Equivalence) The internodons$_A$ are precisely the internodons$_K$.*

Proof. Immediate from Corollary 19. □

Given Theorem 20, it is no longer necessary to distinguish internodons$_A$ from internodons$_K$. Note that Definition 9 is the more algorithmic of the two definitions.

Remark 21. *We have worked in less generality than Kornet et al: for simplicity, we have assumed no two organisms share the exact same birthdate. Kornet et al did not make this assumption. Our work can be generalized to accommodate non-distinct birthdates as follows. If B is a biosphere where some organisms are born simultaneously, let B' be the corresponding (hypothetical) biosphere in which two organisms x and y are identified if x and y have the same birthdate and $x(\boldsymbol{PC}_{\geq x})y$ (note that if x and y are born simultaneously, $x(\boldsymbol{PC}_{\geq x})y$ is equivalent to $y(\boldsymbol{PC}_{\geq y})x$). Then, B' may still contain distinct organisms with shared birthdates, but never one within the dynasty of another. It is straightforward to adjust our work for such a biosphere B'. From the internodons of B', the internodons of B follow by declaring that any x and y identified in B' lie in the same internodon in B.*

Remark 22. *The vigilant reader may have noticed an apparently arbitrary decision (anathema to mathematics) in our definition of pre-internodons. We start by computing undirparents; then, we discard non-***SD** *organisms; then, we take maximal non-branching undirpaths. The first two steps could be reversed: discard non-***SD** *organisms first, compute undirparents second. Assuming no shared birthdates, this would lead to alternate internodon-like clusters. Remark 21 justifies the order we chose: it would not hold if we chose the other order.*

7 The SD Property Revisited

In order to avoid a proliferation of "tiny" permanent splits, Kornet proposed to group non-**SD** organisms with their nearest **SD** ancestors (see also Section 4). This correction causes the internodal tree from Figure 13 to differ from the one in Figure 12b. How great is this difference? One theme of the paper is that the two trees are similar enough to motivate one another, so we feel obliged to demonstrate their similarity: we will argue their difference is negligible, in a formal sense.

At the same time, not all organisms lie in an internodon at all. If an organism is non-**SD** and has no test organism, it is left out of the internodon construction. We will demonstrate that such stray individuals are also negligible.

Unfortunately, the predicate *negligible* is vague. To remove this vagueness, we assume the genealogical network is infinite and treat "negligible" and "small" as synonyms for *finite*. Maybe the genealogical network is *not* infinite! But we would not be the first scientists to approximate the finite by the infinite. See [1] for more justification for this infinitary assumption. Hereafter, we assume the genealogical network is infinite, and we define *negligible* to mean *finite*.

The key to proving negligibility is the Knight–Darwin law, stated by Darwin [2] (see also [3], [1]). Darwin's original words were:

> "...it is a general law of nature that no organic being self-fertilizes itself for a perpetuity of generations; but that a cross with another individual is occasionally—perhaps at long intervals of time—indispensable."

Based on this and surrounding passages, we formalize the Knight–Darwin law as follows (we phrase it as a property of graphs, rather than a law of physics, to emphasize that it is an unfalsifiable inductive generalization).

Definition 13. *(See Figure 16) We say G has the* Knight–Darwin property *if G does not contain any infinite directed path in which every vertex has < 2 parents.*

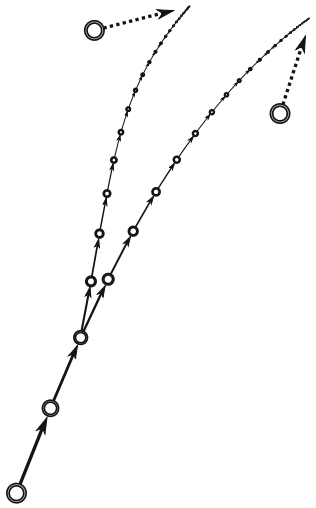

Figure 16: The Knight–Darwin law: every infinite directed path through the genealogical network hits a sexually produced vertex.

We shall need a combinatorial theorem known as König's Lemma.

Theorem 23. *(König's Lemma) Let T be an infinite tree in which every vertex has at most finitely many children. Then T has an infinite directed path.*

Hereafter, we assume that every organism in the biosphere has only finitely many children (we consider this to be a mild assumption).

Theorem 24. *Assume (following [4]) that there are only finitely many parentless organisms. If G has the Knight–Darwin property, then the set $\mathbf{WN}^c$ of non-***SD** *organisms lacking test organisms is negligible.*

Proof. Let $\mathcal{C}$ be the set of connected components of $\mathbf{WN}^c$.

Claim 1: For each $C \in \mathcal{C}$, C contains a parentless organism.

Since connected components are nonempty, C contains some organism x. If x is not parentless, let x' be a parent of x. Then, $x' \in C$ because x' is connected to x, and x' is not **SD** nor does x' have a test organism (in either case, x would have a test organism, contradicting $x \in \mathbf{WN}^c$). If x' is not parentless, let x'' be a parent of x', by identical reasoning $x'' \in C$. This process cannot continue

forever or x would have infinitely many ancestors. Thus, C has a parentless organism, proving the Claim.

It follows (since there are only finitely many parentless organisms) that $|\mathcal{C}| < \infty$.

Claim 2: Every $C \in \mathcal{C}$ is finite.

Let $C \in \mathcal{C}$. By Corollary 10, C is a forest; since C is connected, C is a tree. Assume, for sake of contradiction, C is infinite. By König's Lemma, C has an infinite path. Since G has the Knight–Darwin property, this infinite path has an organism with multiple parents—absurd. The Claim is proved.

Consisting of finitely many finite connected components, $\mathbf{WN}^c$ is finite, so negligible. □

The next corollary shows individual non-**SD** pieces of an internodon are negligible.

Corollary 25. *Let I_0 be a pre-internodon, let I be the corresponding internodon, and let $\mathcal{C}$ be the set of connected components of $I \backslash I_0$. Assuming G has the Knight–Darwin property, each $C \in \mathcal{C}$ is negligible.*

Proof. Let $C \in \mathcal{C}$. By Corollary 10, C is a forest; being connected, C is a tree. If C were infinite, C would contain an infinite directed path by König's Lemma. By the Knight–Darwin property, that infinite path would contain an **SD** organism, absurd. □

8 Conclusions and Future Work

On the biological side, we have presented a new definition of internodons, exploiting the simplification obtained by taking the undirparent relation as our primary focus, as opposed to Kornet's path-connectedness based on the undirancestor relation. This resulted in a tree of life construction that zooms down to individual organisms (with individuals as nodes in the tree). This is theoretically satisfying because it diminishes arbitrary dependence on scale. For example, if we treated individual cells like distinct organisms, the construction would require no adjustments. A topic for future research is to investigate whether the multi-level tree of life can be zoomed all the way down to the level of historical gene trees.

A further spin-off obtained is that the organismal tree generated is homeomorphic to the phylogenetic Hennigian species tree of life. This implies the discovery of a multi-level species tree of life: such a phylogenetic tree can be obtained by zooming out from the organismal tree, or conversely, the organismal tree of life can be generated by expanding the phylogenetic nodes into unary trees.

Lastly, given a birthdated population, with tokogenetic information and data about permanency of splits, an internodon tree can be generated by an efficient algorithm. The latter will be presented in a separate paper.

On the mathematical side, we have taken a DAG along with distinct birthdates and produced a forest. Distinct birthdates are a special case of the well-studied notion of a topological order [6] on a DAG. Given a DAG, the undircon-struction can be viewed as a function taking topological orders and producing forests. We suspect that in some sense, biological undirforests do not depend drastically on which topological order is used; one possible direction of future research would consist of investigating this conjecture.

Acknowledgments

We thank Rino Zandee for inspiring comments and profitable discussions. We thank the anonymous reviewers and the handling editor for their constructive remarks that helped us improve the readability of the paper.

References

[1] Samuel A Alexander. Infinite graphs in systematic biology, with an application to the species problem. *Acta biotheoretica*, 61(2):181–201, 2013.

[2] Charles Darwin. *On the origin of species*. John Murray, 6th edition, 1872.

[3] Francis Darwin. The Knight-Darwin law. *Nature*, 58(1513):630–632, 1898.

[4] Andreas Dress, Vincent Moulton, Mike Steel, and Taoyang Wu. Species, clusters and the 'tree of life': a graph-theoretic perspective. *Journal of Theoretical Biology*, 265(4):535–542, 2010.

[5] Willi Hennig. *Phylogenetic systematics*. University of Illinois Press, 1966 (reprinted 1979).

[6] Arthur B Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

[7] DJ Kornet. Permanent splits as speciation events: a formal reconstruction of the internodal species concept. *Journal of Theoretical Biology*, 164(4):407–435, 1993.

[8] DJ Kornet and James W McAllister. The composite species concept: a rigorous basis for cladistic practice. In *Current themes in theoretical biology*, pages 95–127. Springer, 2005.

[9] DJ Kornet, JAJ Metz, and HAJM Schellinx. Internodons as equivalence classes in genealogical networks: building-blocks for a rigorous species concept. *Journal of Mathematical Biology*, 34(1):110–122, 1995.

[10] Jeremy Martin, David Blackburn, and Edward O Wiley. Are node-based and stem-based clades equivalent? insights from graph theory. *PLoS currents*, 2, 2010.

[11] Kevin C Nixon and Quentin D Wheeler. An amplification of the phylogenetic species concept. *Cladistics*, 6(3):211–223, 1990.

[12] János Podani. Tree thinking, time and topology: comments on the interpretation of tree diagrams in evolutionary/phylogenetic systematics. *Cladistics*, 29(3):315–327, 2013.

[13] Joel D Velasco. The internodal species concept: a response to 'the tree, the network, and the species'. *Biological Journal of the Linnean Society*, 93(4):865–869, 2008.

[14] Joseph H Woodger. From biology to mathematics. *The British Journal for the Philosophy of Science*, 3(9):1–21, 1952.