Journal of Classification 3:299-317 (1986)

Journal of
Classification
©1986 Springer-Verlag New York Inc

N-Trees as Nestings: Complexity, Similarity, and Consensus

Edward N. Adams III

Adams Software

Abstract: Interpreting a taxonomic tree as a set of objects leads to natural measures of complexity and similarity, and sets natural lower bounds on a consensus tree Interpretations differing as to the kind of objects constituting a tree lead to different measures and consensus Subset nesting is preferred over the clusters (strict consensus) and even the triads interpretations because of its superior expression of shared structure Algorithms for computing the complexity and similarity of trees, as well as a consensus index onto [0,1], are presented for this interpretation The "full consensus" is defined as the only tree which includes all the nestings shared in a profile of rival trees and whose clusters reflect only nestings shared in the profile The full consensus is proved to exist uniquely for each profile, and to equal the Adams consensus

Keywords: Full consensus; Adams consensus, Adams-2 consensus; Strict consensus; Rooted trees

1. Introduction

What information is represented by a taxonomic tree? This question is central to discussions of complexity, similarity measures, and consensus of trees, and has had a variety of explicit or implicit answers After an initial period in which they studied many techniques and measures, each with its own rationale, taxonomic researchers have been moving toward the recognition that such measures and techniques should be grouped according to common "interpretation," that is, their assumptions about the nature of information to be found in taxonomic trees However, not all of these underlying interpretations have been presented, and some are inadequate

The author is grateful for the many helpful comments on presentation from Frances McA Adams, William H E Day, and Christopher A Meacham

Author's Address: Edward N Adams III, Adams Software, 2782 Waverley Street, Palo Alto, CA 94306, USA

The most straightforward kind of interpretation is one of the form "a tree is a set of identifiable objects." An example in the realm of fully labeled rooted trees is exemplified in the first consensus algorithm of Adams (1972), where a tree is assumed to be a set of ordered pairs of ancestor-descendant relationships. Such a tree is appropriate for representing evolutionary relationships among OTUs both ancient and modern Any interpretation of a tree as a set of objects (in this case, ordered pairs) is open to simple measures of complexity (the number of objects in the tree) and similarity (the number of objects shared by all the rival trees in a profile), and suggests a lower bound for a consensus tree (the set of objects shared by rivals). A simple virtue of such an interpretation is that the elements of classification are themselves the basis of similarity

This virtue is shared by some interpretations of trees with unlabeled internal vertices Such trees arise from algorithms used to represent taxonomy and to infer phylogeny. One interpretation of such trees as sets of identifiable objects is the clusters interpretation, with its "strict consensus," proposed by, among others, Margush and McMorris (1981), and Sokal and Rohlf (1981). In this interpretation, each branch of a tree is associated with the set of leaves that it separates from the rest of the tree. Each such set is called a *cluster* of the tree, as is the entire set of leaves. A tree is simply a set of clusters In this interpretation, as above, two trees are more similar if they have more elements (clusters, in this case) in common.

Other interpretations rely upon a more abstract, or less immediate, relationship of classification to similarity For example, Robinson (1971) has suggested a similarity measure for unrooted trees based upon a count of nearest-neighbor interchanges. Such an interpretation makes it hard to point to any part of one tree and say "this is what makes this tree similar to that tree." Further, it is hard to characterize the meaning of classification in a tree when transposition of a pair of branches each bearing 20 leaves yields a very similar classification (i e , only one interchange), while transposition of three pairs of branches each bearing one leaf yields a much less similar classification (three interchanges). These advantages may not make it a bad measure, but they do clarify the advantages of an interpretation which more concretely relates similarity to classification.

The interpretation of a tree as a set of clusters has proven a fruitful area of study, yielding a consensus (the strict consensus), a median (the "majority-rule consensus" of Margush and McMorris (1981)), a poset of trees closed under intersection, a variety of metrics of similarity or dissimilarity (see Day (1985) for examples and references), and natural extensions to unrooted trees However, its use as a basis of taxonomic comparison must be due to the mathematical consistency of all these concepts, because it is unable to capture much of the structure that taxonomists intuitively find in trees.

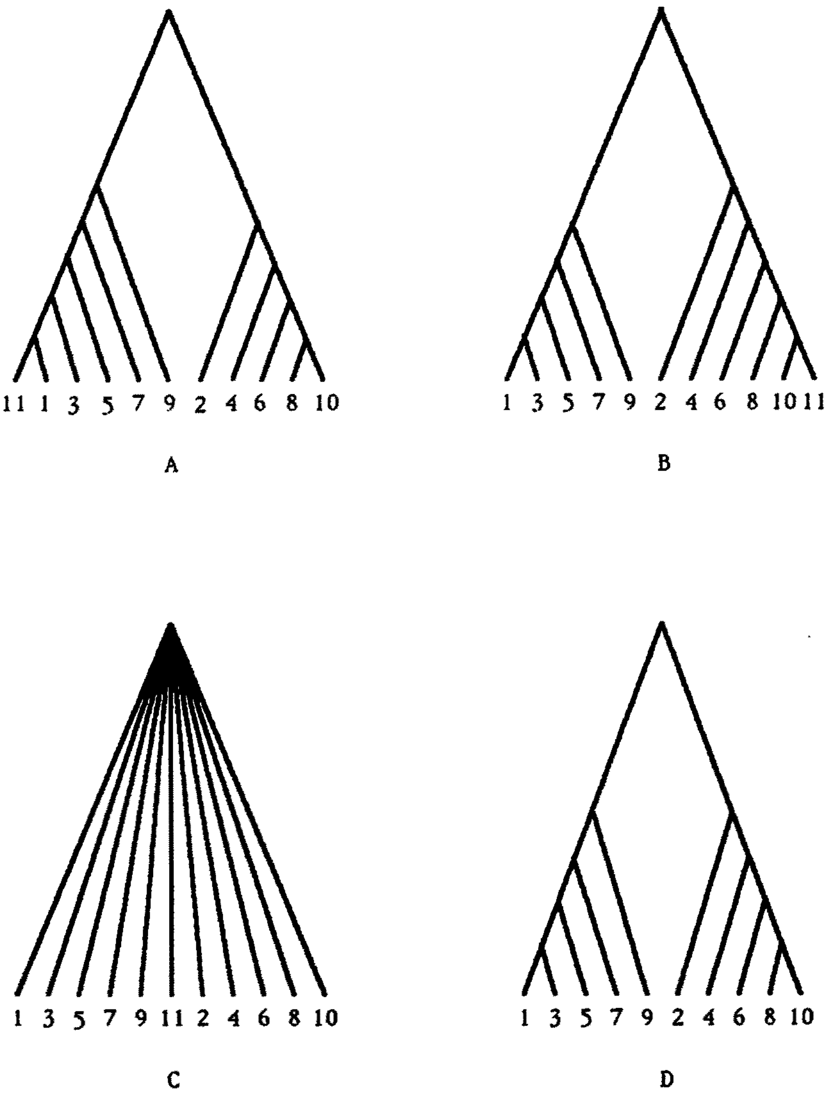

Figure 1 Though the strict consensus of trees A and B is C, the structure common to A and B is better shown in D A cluster is the entire set of leaves below an unlabeled vertex

Figure 1 demonstrates the limitations of the clusters interpretation and the strict consensus, as far as taxonomic structure is concerned. The strict consensus of the two trees A and B is the "bush," or null tree (tree C). Most taxonomists would detect much more shared structure than that, trees A and B seem to share a rather complex tree involving leaves 1-10 (tree D), disagreeing only as to the placement of the leaf 11. But, because no cluster in A equals any in B, the strict consensus finds no shared structure. This is not a fault of the algorithm which could be patched up; the algorithm simply intersects the sets and finds a null intersection. Any extension to the strict consensus that depends on the same interpretation will also be necessarily limited in its ability to capture fine structural detail

A procedure that might have a related interpretation is the *s*-consensus of Stinebrickner (1984), which captures more detail by accepting the intersection of sufficiently similar clusters from the profile. In this way, it can extract clusters even when they are not attested exactly in any one rival tree. This implies that a tree is interpreted as a set of something other than clusters. But without knowing what kind of information a tree contains, we cannot know how much the *s*-consensus preserves and how much it discards Thus, it is hard to know how the *s*-consensus index relates to that shared information

The same problems weaken our understanding of any method whose underlying interpretation is not known, such as my original consensus algorithm from Adams (1972), now called "Adams consensus" or "Adams-2 consensus." We should identify and evaluate the interpretations underlying existing methods, in order to expose inherent strengths and weaknesses In new domains, we can similarly evaluate candidate interpretations, even before investing the effort to create algorithms or measures.

So, in searching for a newer and better interpretation, consider the proposal "a tree is a set of triads" If two leaves, *a* and *b*, separate further from the root than they do from a third leaf, *c*, then we can say that *a* and *b* *discriminate against* *c*, or the triad $t(a,b,c) = c$, whereas if they all separate at the same place, we can say that $t(a,b,c) = 0$, or no triad exists for (a,b,c). The concept of triads for rooted trees is analogous to the "quartet" concept for unrooted trees proposed by Estabrook, McMorris, and Meacham (1985). Under this interpretation of a tree as a set of triads, the intersection over a profile is the set of triads shared by all the rivals in the profile. Although this intersection does not always describe a tree on the entire set of leaves, the incomplete tree(s) that it constitutes can be quite suggestive. For example, tree D from Figure 1 embodies all and only the triads shared by trees A and B. Of course, wherever leaf 11 is added to this tree, additional triads will be implied which were *not* shared in the profile A rationale for choosing a place for leaf 11 could be devised, yielding a definition for a consensus tree, and the complexity and similarity measures could be based on the number of triads and the number of shared triads, respectively.

But as promising as this interpretation is, it still fails to capture all the structural information of a tree, as shown in Figure 2. Disagreeing with Margush and McMorris (1981), I observe that in rivals A and B, leaves 1 and 2 join at a lower level (i.e., further from the root) than the whose set does, and I believe that that closer relationship (which I will call *subset*

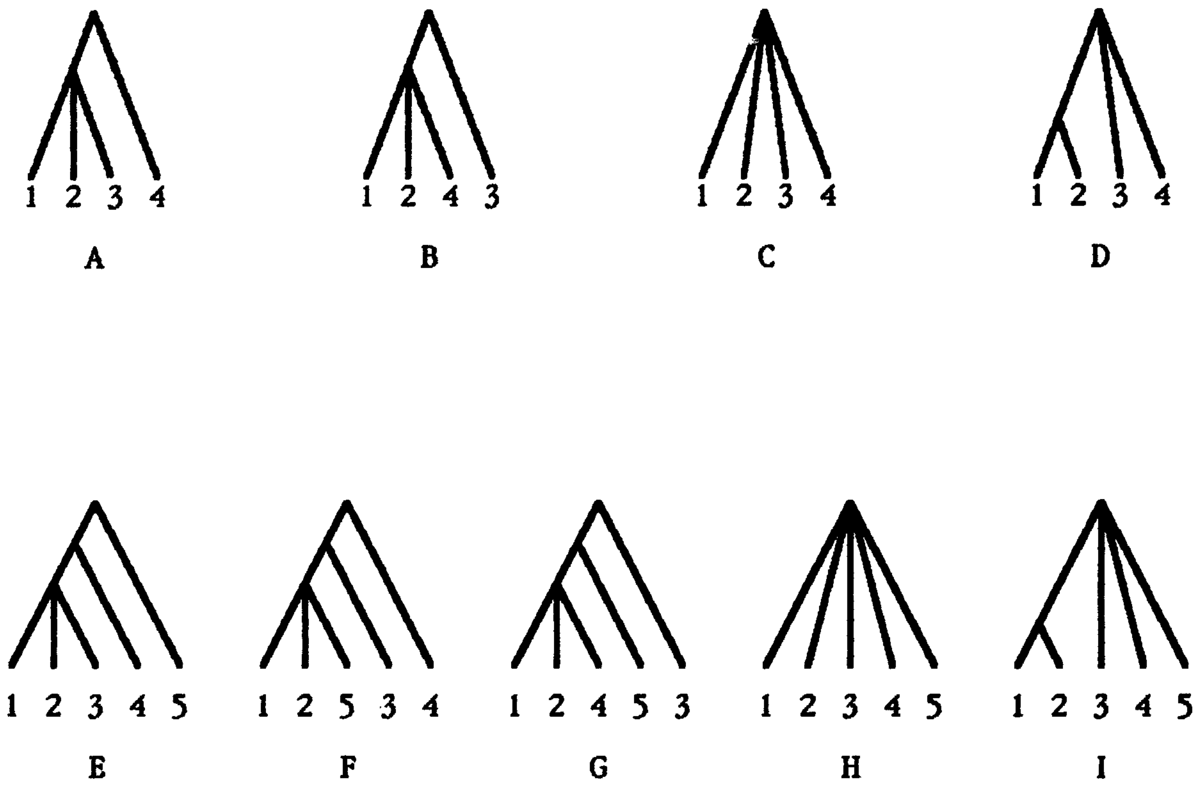

Figure 2. Even the triad interpretation misses some structure. Trees C and D are the triad consensus and full consensus of A and B; H and I are the triad consensus and full consensus of E, F, and G.

nesting) is part of the information represented by both trees. As the trees share no triads, however, the intersection of the triads yields tree C, the bush. Possibly more compelling is the case of rivals E, F and G, where leaf subset {12} is nested in a larger set {12345}. The consistency of 1 and 2 deserves more recognition than they get in the triad intersection H. Thus, the interpretation of a tree as a set of triads is unsatisfactory. Extending from triads to pentads would still fail due to analogous counterexamples.

A more extreme interpretation, "a tree is a set of *leaf subset nestings*," permits the expression of this closer relationship of leaves 1 and 2. As it defines a tree as a set of identifiable objects, its complexity and similarity measures are simple counts. The shared information, exemplified in trees D and I, retains much more detail than the strict consensus, and even more than the triads consensus. In the remainder of this paper, I define the nestings interpretation, show how to compute its complexity and similarity measures, define its ("full") consensus as the best tree that represents all the information shared among a set of trees, and prove that the Adams consensus is the full consensus.

2. Leaf Subset Nesting

The different definitions of the word *tree* used in this paper are equivalent mathematically, but some reveal taxonomic structure more effectively than others.

A graph-theoretic definition of *tree* is an undirected acyclic connected graph with one distinguished vertex (the *root)* of degree greater than 1, an optional set of unlabeled vertices of degree greater than 2, and n uniquely labeled vertices of degree 1 (the *leaves)* In the illustrations, a tree is depicted with its root at the top and the leaves at the bottom. Similarly, the text uses the words *up* and *down* to mean "toward the root" and "away from the root," respectively.

Another definition of *tree* is what Margush and McMorris (1981) call an "n-tree" a set T of subsets of $N(=\{1, \ldots, n\})$ satisfying the conditions

$$N \in T\,, \tag{C1}$$

$$\varnothing \notin T\,, \tag{C2}$$

$$\{i\} \in T \textit{ for every } i \textit{ in } N\,, \tag{C3}$$

$$X \cap Y \quad \{\varnothing, X, Y\} \textit{ for every } X \textit{ and } Y \textit{ in } T\,. \tag{C4}$$

The universe of such trees is called R_n. The nonsingleton elements of a set T are called the *clusters* of the tree.

There is a natural bijection between graph-theoretic trees and n-trees. Each leaf vertex corresponds to a singleton element of T. Each nonleaf vertex is the root of a subtree on a subset of the leaves, and corresponds to the cluster of T equal to that subset of leaves. Condition (C4) corresponds to the acyclicity of the graph, so that there is only one path from the root to a given vertex. As an example of this bijection, tree E of Figure 2, viewed as a graph, consists of 8 vertices, three of which are unlabeled (the top one being the root). As an n-tree, or set of clusters, it consists of the 8 sets $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{123\}$, $\{1234\}$, and $\{12345\}$

If $X, Y \in T$, and $X \subset Y$, then we say that $X < Y$. The *least upper bound* in tree T of any subset M of N is the smallest $X \in T$ such that $M \subseteq X$, and is denoted $lub_T(M)$. If $X < Y$, we call X a *descendant* of Y, and Y an *ancestor* of X, and if there is no $Z \in T$ such that $X < Z < Y$, then we call X a *child* of Y, and Y the *parent* of X We also use the ancestry terminology when referring to the corresponding vertices of the tree viewed as a graph

The previous section introduced the notion of a *nesting* — one set of leaves joining lower than an including set We can use this notion as the basis of a third interpretation of the concept of tree Consider $<_T$ to be a relation on the Cartesian product *Power*^{2} (N), where *Power* (N) is the set of all subsets of N, so that $\forall\ i \in N, \forall$ nonnull $A, B, C \subset N$,

$$A <_T B \rightarrow A \subset B\ , \tag{C5}$$

$$A \neq \{i\} \rightarrow \{i\} <_T A \cup \{i\}\ , \tag{C6}$$

$$\begin{aligned} A <_T B \rightarrow \Big[& (C \subset A \rightarrow C <_T B) \wedge \\ & (A \subset C \subset B \rightarrow A <_T C \vee C <_T B) \wedge \\ & (B \subset C \rightarrow A <_T C)\Big]\ , \end{aligned} \tag{C7}$$

$$A <_T C \wedge B <_T C \rightarrow (A \cup B <_T C \vee A \cup B = \emptyset)\ . \tag{C8}$$

Any ordered pair (A,B) for which the relation holds is called a *nesting*, and we say that A *nests in* B, or B *houses* A. Condition (C5) asserts that only a strict subset can nest. Condition (C6) asserts that any singleton nests in all larger supersets. Condition (C7) expresses a limited form of transitivity, such that if we have a nesting (A,B), any set included in A nests in anything A does, any set strictly between A and B must be in a nesting with at least one of them, and any set that includes B houses A. Condition (C8) is analogous to (C4), in that if one set houses two sets, then either it also houses their union, or they are disjoint.

Each such relation $<_T$ corresponds to a tree T in R_n by the following correspondence. $X <_T Y$ iff $X \subset Y \wedge lub_T(X) < lub_T(Y)$. To construct a tree T in R_n from a relation $<_T$, construct the cluster $L(S)$ for each non-null $S \subseteq N$ to be $\{c \in N \mid \neg(S <_T S \cup \{c\})\}$. Then T is the union of the set of distinct clusters $\{L(S) . \varnothing \subset S \subseteq N\}$ and the singletons. Theorem 1 in Appendix A proves that the bijection holds. To understand this construction, see that $L(S)$ is the largest superset of S that S does not nest in. Viewed in the n-tree, $L(S)$ is simply $lub_T(S)$. For example, in tree E of Figure 2, where $\{12\} < \{124\}$, $\{12\} < \{125\}$, $\neg(\{12\} < \{123\})$, we have $L(\{12\}) = \{123\}$. Similarly, $L(\{13\}) = \{123\}$ and $L(\{14\}) = \{1234\}$

Whenever a classification is defined to be a set of objects, natural definitions of complexity, similarity, and consensus follow. Using the nestings interpretation of a tree, define the *nesting complexity* $c_n(T)$ of a tree T to be the number of its nontrivial nestings (i.e., nestings not also present in the null tree). This is analogous to the *component information* of an n-tree, as defined by Nelson (1979).

The consensus over a profile of sets is their intersection, so define the *intersection* $\cap_P$ of a profile P of trees to be a relation $\mathbf{R}$ such that $\forall A,B \subset N$, $A \mathbf{R} B$ iff $(\forall p \in P)\ A <_p B$. In other words, construct the relation consisting of all the nestings that are in all the rivals of the profile. Because it is the intersection, it is analogous to the strict consensus of n-trees.

Finally, the similarity over a profile of sets is the size of their intersection, so define the *nesting similarity* $s_n(P)$ to be the nesting complexity of the intersection It is analogous to the component information of the strict consensus.

Figure 3 and Table 1 show examples of 5-trees and their nontrivial nestings, respectively. The algorithms for computing s_n and c_n are in Appendix B.

From this point of view, the perceived similarity of trees A and B of Figure 1 can be quantified ($c_n(A) = c_n(B) = 23129$ and $s_n(A,B) = 8204$), as can the failure of the strict consensus to capture it ($c_n(C) = 0$), but it is hard to relate this result either to other definitions of consensus or to comparison of other sized trees. To facilitate comparisons of comparisons, Day (1983) suggests that a consensus index should range from 0 (no similarity) to 1 (unanimity) inclusive, based upon a complexity measure that is an interval scale. An interval scale can be based upon logarithmic transformations of c_n and s_n, as the asymptotic behavior of c_n and s_n is exponential. It is important to have a sensible value for the null tree, so, to avoid $ln(0)$, we establish the transformations $lc_n(t) = ln(1 + c_n(t))$ and $ls_n(P) = ln(1 + s_n(P))$. No logarithm base is to be preferred over another

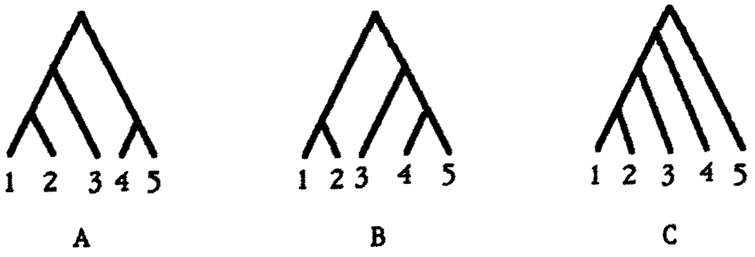

Figure 3 Three 5-trees

Table 1

Nestings for Figure 3 [a]

Tree A		Tree B		Tree C	
(12)3	(12)34	(12)3	(12)34	(12)3	(12)34
(12)345	(12)35	(12)345	(12)35	(12)345	(12)35
(12)4	(12)45	(12)4	(12)45	(12)4	(12)45
(12)5	(123)4	(12)5	(34)1	(12)5	(123)4
(123)45	(123)5	(34)12	(34)125	(123)45	(123)5
(13)24	(13)245	(34)15	(34)2	(1234)5	(124)35
(13)25	(13)4	(34)25	(345)1	(13)24	(13)245
(13)45	(13)5	(345)12	(345)2	(13)25	(13)4
(23)14	(23)145	(35)1	(35)12	(13)45	(13)5
(23)15	(23)4	(35)124	(35)14	(134)25	(134)5
(23)45	(23)5	(35)2	(35)24	(14)235	(14)25
(45)1	(45)12	(45)1	(45)12	(14)35	(14)5
(45)123	(45)13	(45)123	(45)13	(23)14	(23)145
(45)2	(45)23	(45)2	(45)23	(23)15	(23)4
(45)3		(45)3		(23)45	(23)5
				(234)15	(234)5
				(24)135	(24)15
				(24)35	(24)5
				(34)125	(34)15
				(34)25	(34)5

[a] (X)Y means $X < X \cup Y$

for this scale, because we will ultimately be interested in ratios of complexity values. In this scale, one interval represents the difference in complexity between a tree of any size n and a similar one of size $n + 1$. Taking Figure 1 as an example, $lc_n(A) = 10.04$ and $ls_n(\{A,B\}) = 9.01$. These numbers convey the impressionistic notion that tree A has as much nesting complexity as an idealized tree with 10.04 leaves, while the complexity of its intersection with B is that of an idealized tree of only 9.01 leaves. In other words, the scale describes the size (number of leaves) of an equivalent idealized tree. (Of course, the "number of leaves" is impressionistic, because the logarithm base is arbitrary.)

Day suggests estimating the fit of a nominal consensus to a profile by the ratio of the complexities of the consensus and the profile. Following the spirit rather than the letter of his prescription, I interpret profile complexity as consensus complexity (i.e., nesting complexity of the consensus) plus dispersion (defined below) about the consensus. The complexity due to information unique to a tree t compared to profile P is $lc_n(t) - ls_n(P \cup \{t\})$, while that due to information unique to profile P is $ls_n(P) - ls_n(P \cup \{t\})$, so define $d(t, P)$, the *deviation* of a tree t from a profile P, to be their sum

$$d(t, P) = lc_n(t) + ls_n(P) - 2ls_n(P \cup \{t\}) .$$

Notice that if P contains t, this reduces to $lc_n(t) - ls_n(P)$, while if P just has one member u, it reduces to $lc_n(t) + lc_n(u) - 2ls_n(\{t,u\})$. In the nestings domain, the role of consensus can be played by either a tree or an intersection of trees (over either the profile of interest or some other profile), as intersections show the shared complexity exactly. So if the intersection R_Q over a profile Q is proposed as a nominal consensus for profile P, define the *dispersion* of P about R_Q as the deviation of the rivals of P from Q, averaged among the rivals, $|P|^{-1} \sum_{p \in P} d(p,Q)$. Following Day (in spirit), we can define the *nesting fit* of R_Q to P as the ratio of consensus complexity to the sum of consensus complexity and dispersion about the consensus

$$f_n(Q,P) = ls_n(Q) / \ (ls_n(Q) + |P|^{-1} \textstyle\sum_{p \in P} d(p,Q)) .$$

If the denominator vanishes, define $f_n(Q,P)$ to be 0 This only occurs if all of the rivals in P and the intersection over Q are null trees.

This measure evaluates the intersection over profile Q as a consensus for profile P. Because Q can consist of as little as one tree, the measure is useful for evaluating consensus trees produced by any candidate consensus method. We define the special case where Q and P are equal as the *nesting* *consensus index* $ci_n(P) \equiv f_n(P,P)$. It can easily be shown that

$$ci_n(P) = |P|\ ls_n(P)/ \textstyle\sum_{p \in P} lc_n(p)\ .$$

This index qualifies as a consensus index because it achieves unity exactly when P is unanimous and contains no null trees, tends toward zero if the dispersion increases, and is zero if $ls_n(P)$ is To complete the example from Figure 1, $ci_n(\{A,B\}) = (2 \times 9.01)/\ (2 \times 10\ 04) = 0\ 892$.

3. Full Consensus

For statistical purposes, the comparison process can stop right here. the nestings of the intersection constitute all the shared information, and their number is known. If, however, we wish to visualize the shared information, we need to build a tree from the intersection. We shall call such a tree the "full" consensus of a profile, as it should include *all* of the nesting information shared by all the rival trees in the profile An obvious definition of the full consensus tree C is that it should satisfy conditions (R1) and (R2).

$$(\forall\ A,B \subseteq N)\ ((\forall\ T \in P)\ A <_T B) \rightarrow A <_C B \qquad \text{(R1)}$$

$$(\forall\ A,B \subseteq N)\ A <_C B \rightarrow ((\forall\ T \in P)\ A <_T B) \qquad \text{(R2)}$$

In other words, all the nestings common to the profile should be in C, and all the nestings in C should be common to the profile Unfortunately, these conditions are too strong, as exemplified by Figure 4. The only shared nestings are $\{23\} < \{234\}$ and $\{23\} < \{1234\}$, which justify the existence of two clusters $\{23\}$ and $\{234\}$, but there is no place for 1, because neither $\{23\} < \{123\}$ nor $\{123\} < \{1234\}$ is in the intersection.

Instead of requiring that all nestings of the consensus reflect nestings of all the rival trees, we could merely require that the most obvious nestings do so Thus condition (R2') requires only that the clusters of C reflect nestings shared by all

$$(\forall\ A,B \in C)\ A <_C B \rightarrow ((\forall\ T \in P)\ A <_T B) \qquad \text{(R2')}$$

Although this seems too weak, Theorem 2 (in Appendix A) shows that there is no more than one tree satisfying (R1) and (R2') for a profile P. In fact, the full consensus tree exists for each profile and turns out to be the Adams consensus tree, as shown in Theorem 3 (in Appendix A) Although $s_n(P) \leqslant c_n(C)$ for the full consensus C of P, they are equal for the cases of unanimity and total disagreement, so the full consensus and consensus index ci_n constitute a CI method, as defined in Day and McMorris (1985).

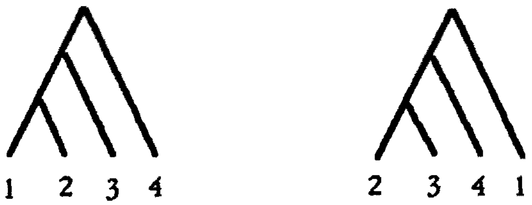

Figure 4 The intersection of trees can fail to be a tree The shared nestings are (23)4 and (23)14 Either (23)1 or (123)4 is also needed to satisfy the tree conditions in the nestings domain

Some drawbacks of the full consensus tree have been noted it is uncharacterized, its computation is moderately expensive, and it can contain clusters not present in any rival The first objection was an impetus for writing this paper. The second is also valid, but computing power is getting less expensive each year. The third objection carries weight insofar as one feels that a tree is a set of clusters. But Figure 1 shows that the structure of a tree is far more complex than is manifest in a set of clusters. The interpretation of a tree as a nesting relation between sets squeezes every drop out of the notion of structural information of a tree, and the full consensus tree merely represents that shared information. Although it is true that it usually also represents additional information, the "uninterpretable" clusters *do* have an interpretation, by Theorem 3. any nondisjoint pair of clusters of the full consensus are in fact nested (i e., obey the $<$ relation) in all the rivals. Theorem 2 says that no other consensus tree can make this claim (and still portray all the shared nestings)!

4. Conclusion

Understanding how a tree is interpreted to display classificatory information is important for evaluating existing complexity and similarity measures and consensus procedures. It can also help direct searches for new ones, especially in other domains, such as unrooted trees and weighted trees.

This paper has examined three interpretations of a tree as a set — clusters, triads, and nestings — and shown them to be in increasing order of inherent structural complexity. For the nestings interpretation, it has defined raw complexity and similarity measures, and normalized consensus index and consensus fit measures analogous to one proposed by Day (1983) Finally, it has defined the "full consensus" tree while guaranteeing that its nondisjoint clusters bear the nesting relation in each tree of the profile, and has shown that the full consensus is the same as the Adams consensus.

Appendix A

Here are some useful lemmas in the domain of subset nestings.

Lemma 1 $A \cup B < B \cup C \wedge A \subset C \rightarrow A < C$

Proof Assume $A \cup B < B \cup C$ and $A \subset C$. Then (C7) implies $A < B \cup C$ Therefore, if $\neg(A < C)$ then, by (C7) But also $C < B \cup C$, so, by (C8), $A \cup B \cup C < B$ or $C \cap (A \cup B) = \varnothing$, both of which are false So $A < C$ ●

Lemma 2 $A \subset B \wedge A \subset C \wedge \neg(A < B \vee A < C) \rightarrow \neg(A < B \cup C)$

Proof Assume $A \subset B$, $A \subset C$, and $\neg(A < B \vee A < C)$. Suppose $A < B \cup C$. Then, by (C7), $\neg(A < B)$ implies $B < B \cup C$ and $\neg(A < C)$ implies $C < B \cup C$ These, with (C8), imply $(B \cup C < B \cup C)$ $(B \cap C = \varnothing)$, both of which are false, so $\neg(A < B \cup C)$. ●

Theorem 1 *There is a bijection between R_n and V_n, the set of objects that satisfy the conditions (C5)-(C8).*

Proof The proof proceeds in two parts

From T in R_n to V in V_n Construct the function p $R_n \rightarrow V_n$ as follows If X is a cluster of T, declare that $S <_V S' \cup R$, $\forall$ nonnull $S \subseteq S' \subseteq X$ and $\forall$ nonnull $R \subseteq N - X$ (C5) is satisfied by definition, and (C6) is satisfied because $\forall$ $i \in N$, $\{i\} \in T$. The first and third parts of (C7) are satisfied by definition If $S < W$ is a nesting that results from this construction based upon X, then for all S' between S and W that are subsets of X, $S' <_V W$ also results, while for all S' between S and W that are not subsets of X, $S <_V S'$ results, so (C7) is satisfied If $A <_V C$ and $B <_V C$ have been constructed, then there were two clusters X_1 and X_2 (possibly equal) such that $A \subseteq X_1$, $B \subseteq X_2$, $C - X_1 \neq \varnothing$, and $C - X_2 \neq \varnothing$. Since $X_1 \cap X_2 \in \{\varnothing, X_1, X_2\}$, A and B are either disjoint or subsets of the smaller of X_1 and X_2 The clauses of (C8) are satisfied by one of the two cases.

Thus, for each element of R_n there is an element of V_n. For any distinct pair of trees T and U in R_n, there must be a cluster X in one tree (call it T) that is not in the other. By the construction, there are nestings $X < X \cup \{c\}$ in $p(T)$ for all $c \in N - X$, but there is no way for all of those nestings to be constructed in $p(U)$ So distinct elements of R_n are related to distinct elements of V_n

From V in V_n to T in R_n. Construct the function $q.\ V_n \to R_n$ as follows. Construct $L(S) \equiv \{c \in N \mid \neg(S < S \cup \{c\})\}$, for all nonnull $S \subset N$. We need to show that conditions (C1)-(C4) are satisfied.

Conditions (C1) and (C2) are obvious by construction. (C3) follows from (C2). Suppose $\exists X = L(R)$ and $\exists Y = L(S)$ for which $X \cap Y \notin \{\varnothing, X, Y\}$. Then $\exists e \in X - Y \wedge \exists f \in Y - X$. By definition of L, $\neg(R < R \cup \{e\})$, $\neg(S < S \cup \{f\})$, $R < R \cup \{f\}$, and $S < S \cup \{e\}$. By Lemmas 1 and 2, $\neg(L(R) < L(R) \cup \{e\})$, $\neg(L(S) < L(S) \cup \{f\})$, $L(R) \cup \{f\}$, and $L(S) < L(S) \cup \{e\}$. Renamed, the last two clauses are $X < X \cup \{f\}$ and $Y < Y \cup \{e\}$. Because $e \in X$ and $f \in Y$, (C7) implies $X < X \cup Y$ and $Y < X \cup Y$ Consequently, (C8) implies $X \cup Y < X \cup Y$ or $X \cap y = \varnothing$, both of which are false. This contradiction shows that all pairs of clusters are either nested or disjoint, so a tree defined in terms of nestings is equivalent to a tree defined in terms of clusters.

To show that the correspondence is a bijection, we need to show that for any V in V_n, $p(q(V)) = V$. Suppose $\exists V_0, V_1 \in V_n$ such that $V_1 = p(q(V_0))$.

For all S, $S' \subseteq N$ such that $S <_0 S'$, then $S <_0 S' \cup L(S)$, by (C7), and as $\neg(S <_0 L(S))$ by repeated applications of Lemma 2, $L(S) <_0 S' \cup L(S)$, by (C7). We can rewrite this as $L(S) <_0 L(S) \cup (S' - L(S))$. Function q created a cluster equal to $L(S)$, and function p created the nesting $S <_1 S'$, because $S \subseteq L(S)$ and $(S' - L(S)) \subseteq N - L(S)$.

On the other hand, consider $\forall\ S \subseteq S' \subseteq N$ such that $\neg(S <_0 S')$. By repeated application of Lemma 2, $\neg(S' <_0 L(S'))$, so $\neg(S <_0 L(S'))$ and $\neg(L(S) <_0 L(S'))$, both by the contrapositive of (C7). Function q produced a cluster equal to $L(S) = L(S')$, so it could not produce a cluster X where $S \subseteq X$ but $S' - X \neq \varnothing$, so function p could not produce $S <_1 S'$, therefore $\neg(S <_1 S')$. Since $(\forall\ S \subseteq S' \subseteq N)\ S <_0 S'$ iff $S <_1 S'$, $V_0 = V_1$. ●

To prove the theorems about the full consensus tree, we need a few definitions. Define $D(x,T)$ to be the sequence $D(x,T)[1], \ldots, D(x,T)[n_x]$ of clusters of tree T that contain the leaf x, in order of decreasing size, so that $D(x,T)[1] = N$ and $D(x,T)[n_x] = \{x\}$. If leaf $x \in S$, and $lub_T(S) = L$, then define $C_{xT}(L)$ to be that child of L in T that contains x. So, for example, $D(x,T)[i+1] = C_{xT}(D(x,T)[i])$.

Theorem 2 *No more than one tree satisfies conditions (R1) and (R2') for a profile P of rival trees.*

Proof If two trees S and T satisfy the conditions, we will show that $D(x,S) = D(x,T)$ for all x, by induction on the elements of D. Clearly $D(x,S)[1] = D(x,T)[1] = N$ Now if $D(x,S)[k] = D(x,T)[k]$, we will refer to this set as $D[k]$. Because S satisfies (R2'), $D(x,S)[k+1] <_S D[k] \rightarrow (\forall\, p \in P)\ D(x,S)[k+1] <_p D[k]$. Then, because T satisfies (R1), $D(x,S)[k+1] <_T D[k]$, and, by (C8), $(D(x,S)[k+1] \cup D(x,T)[k+1]) <_T D[k]$. Call this union U. Suppose $D(x,S)[k+1] - D(x,T)[k+1] \neq \varnothing$. Because $D(x,T)[k+1]$ lacks some member of U, $D(x,T)[k+1] <_T U <_T D[k]$ But this contradicts the fact that $D(x,T)[k+1]$ and $D[k]$ are adjacent in tree T. So $D(x,S)[k+1] - D(x,T)[k+1] = \varnothing$ A symmetric argument shows $D(x,T)[k+1] - D(x,S)[k+1] = \varnothing$, so they are equal and the induction hypothesis is advanced. ●

Theorem 3 *For any profile P, Algorithm 2 in Appendix B computes a tree T_c that satisfies (R1) and (R2') for P.*

Proof To show that T_c satisfies (R1), suppose $\exists\, S \subset S' \subseteq N$ such that $\forall\, p \in P\ (S <_p S')$ For any $x \in S$, find k such that $D(x,T_c)[k] = lub_c(S')$ and call this set S''. Clearly $D(x,T_c)[k+1] \subset S' \subseteq S''$ and $(\forall\, p \in P)\ S' \subseteq lub_p(S'')$. We need to show that $S \subseteq D(x,T_c)[k+1]$, as $S <_c S'$ follows immediately

The algorithm will produce $D(x,T_c)[k+1] = \cap_p\ C_{xp}(lub_p(S'')) \subset S'$, so $\exists\, q \in P$ such that $\neg(S' \subseteq C_{xp}(lib_q(S'')))$, or the intersection would contain S'. For all such q, $lub_q(S') = lub_q(S'')$ by definition of *lub*, and because $S <_q S'$, $lub_q(S) <_q lub_q(S'')$ Thus $lub_q(S) \leqslant_q C_{xq}(lub_q(S''))$ and

$$S \subseteq lub_q(S) \subseteq C_{xq}\ (lub_q(S''))\ . \tag{1}$$

But for those $p \in P$, if any, for which $S' \subseteq C_{xp}(lub_p(S''))$,

$$S \subseteq S' \subseteq C_{xp}\ (lub_p(S''))\ . \tag{2}$$

As (1) and (2) are the only cases, $\forall\, p \in P\ (S \subseteq C_{xp}(lub_p(S'')))$, and S is a subset of their intersection, which is exactly $D(x,T_c)[k+1]$.

To show that T_c satisfies (R2'), assume S, $S' \in T_c$, and $S \subset S'$. We need to show that $(\forall\, p \in P)\ lub_p(S) <_p lub_p(S')$ Use the algorithm to find the sequence $D(x,T_c)$ for some $x \in S$, and identify i and k, where $D(x,T_c)[i] = S$ and $D(x,T_c)[k] = S'$ Then the algorithm computed $D(x,T_c)[k+1] = \cap_p\ C_{xp}(lub_p(S'))$, so $(\forall\, p \in P)\ D(x,T_c)\ [k+1] \subseteq C_{xp}(lub_p(S'))$, which implies

$$(\forall\ p \in P)\ lub_p(D(x,T_c)[k+1] \leqslant_p C_{xp}(lub_p(S')) \tag{3}$$

By combining (3), the assumption that $S = D(x,T_c)[i] \subseteq D(x,T_c)[k+1]$, and an implication of the definition of C_{xp} that $(\forall\ p \in P)$ $C_{xp}(lub_p(S')) < lub_p(S')$, we get $(\forall\ p \in P)$ $lub_p(S) = lub_p(D(x,T_c)[i]) \leqslant_p lub_p(D(x,T_c)[k+1]) \leqslant_p C_{xp}(lub_p(S') <_p lub_p(S')$. The first and last terms of this expression show $(\forall\ p \in P)\ S <_p S'$. ●

Appendix B

Algorithm 1.

Computing c_n If $c(i)$ is the size of cluster i, and $p(i)$ is the size of the parent of cluster i, then

$$c_n(T) = \sum_{i \in T-\{N\}, c(i)>1} (2^{p(i)-c(i)} - 1)\ (3^{c(i)} - 2^{c(i)} - c(i)2^{c(i)-1})\ .$$

The derivation proceeds upward from the leaves The sum consists of contributions from each cluster of the tree T. For purposes of exposition, let us refer to the members of the phrase "$A < B$" as the "twig" A and the "nest" B. Call a cluster of interest S and its parent S'

If $|S| = c$ and $|S'| = p$, then any one particular twig of size i from S is included in $2^{c-1}\ (2^{p-c} - 1)$ nests from the set S' that include at least one element of $S' - S$ There are $C(c,i)$ such twigs (where $C(c,i) \equiv c!/\ i!\,(c-i)!$), and we are interested in twigs of sizes 2 through c, so the contribution due to cluster S is $(2^{p-c} - 1) \sum_{2 \leqslant i \leqslant c} 2^{c-i}\ C(c,i)$. Recognizing that $\sum_{i \leqslant c} 2^{c-i}\ C(c,i) = 3^c$, $C(c,1) = c$ and $C(c,0) = 1$, we can express this as $(2^{p-c} - 1)\ (3^c - 2^c - c2^{c-1})$ To compute $c_n(T)$, add the contributions for all nonsingleton $S \in T-\{N\}$. By counting all twigs included in S, but only nests included in S', we avoid duplication. ●

Because the method for computing s_n mirrors the operation of the Adams consensus algorithm defined in Adams (1972) and more succinctly stated in Neumann (1983), I will restate that algorithm here, after a few definitions. First, if P is a set of disjoint nonempty sets with union U, we call P a *partition* of U. If P is a partition of U, and V is a subset of U, define the *partition of V induced by P* to be the partition Q such that $(\forall\ q \in Q)\ (\exists\, p \in P)\ q = V \cap p$. Further, if PP is a set of partitions P_i of U, define the *partition product* of PP to be the partition Q of U such that $(\forall\ q \in Q)\ (\forall\ P_i \in PP)\ (\exists\, p_i \in P_i)\ q = \cap_i\, p_i$.

Algorithm 2.

Computing the Adams consensus tree T_c over a profile P of trees T_i Start with T_c containing the cluster N. Repeat the following for each nontrivial element S of T_c Let $PP_i(S)$ be the partition of S induced by the children of $lub_i(S)$. Construct the new partition $PP_c(S)$ to be the partition product of the $PP_i(S)$ over all T_i in P Let $T_c = T_c \cup PP_c(S)$ ●

In order to describe the algorithm for computing s_n, we need a supplementary counting formula To compute $G_k(\{X_i\})$, the number of subsets of $\cup_{i \leqslant k} X_i$ that overlap each of the X_i, we develop a recursive formula, and then the more important closed forms Suppose that we know $G_p(\{X_i\})$ for $p < k$. If we add a set, X_{p+1}, we might add some desired subsets Let e ("extra") be $|X_{p+1} - \cup_{i \leqslant p} X_i|$, the number of elements of X_{p+1} that have never been in subsets before Now compose any set counted in $G_p(\{X_i\})$ with any subset of the extra set. There are $2^e\ G_p(\{X_i\})$ of these, nearly all of which should be members of $G_{p+1}(\{X_i\})$ The only composed sets to be rejected are those that don't overlap X_{p+1}. These were composed of (a) an element of $G_p(\{X_i\})$ which did not overlap X_{p+1}, and (b) the null subset of the extra set There are $G_p(\{X_i - X_{p+1}\})$ of these rejects, so the recursive formula is.

$$G_1(\{X_i\}) = 2^a - 1, \textit{ where } a = |X_1|,$$

$$G_{p+1}(\{X_i\}) = 2^e G_p(\{X_i\}) - G_p(\{X_i - X_{p+1}\}), \textit{ where } e = |X_{p+1} - \cup_{i \leqslant p} X_i|$$

The closed forms of these formulas for $k = 2$ and 3 follow Let $a = |X_1|$, $b = |X_2|$, and $c = |X_3|$, and let $ab = |X_1 \cap X_2|$, etc As these literals are never multiplied, this will not be ambiguous.

$$G_2(\{X_i\}) = 2^{-ab}(2^{a+b} - 2^a - 2^b) + 1\ .$$

$$G_3(\{X_i\}) = 2^{abc-ab-ac-bc}\ (2^{a+b+c} - 2^{a+b} - 2^{a+c}$$

$$- 2^{b+c} + 2^{a+bc} + 2^{b+ac} + 2^{c+ab}) + 1\ . \quad ●$$

Algorithm 3.

Computing s_n over a profile P of trees T_i, $2 \leqslant i \leqslant |P|$ For each cluster S of T_c other than the root, consider a twig that joins at S and all nests that include it and are shared by the trees in P

Counting the twigs is easy To avoid multiple counting, we only count a twig that joins at S, namely one that contains at least one element from more than one "child" (member of $PP_c(S)$). This is simply the number of twigs of size s from S, less the number of twigs of size s that derive from only one child, $C(|S|,s) - \sum_j C(|PP_c(S)[j]|,s)$, where i ranges over members of $PP_c(S)$ at least as large as s.

We count all nests that include the twig and certain sets of outliers. An outlier in T_i is any of the elements of $N - lub_i(S)$. For a nest to be shared by the trees in P, it must contain at least one outlier from each tree. Then the number of nests around a twig of size s is $2^{m-s}\ G_{|P|}\ (\{N - lub_i(S)\})$, where $m = |S|$. Multiplying by the number of twigs, we find the contribution fro a given set S to be

$$G_{|P|}(\{N - lub_i(S)\})\sum_{2\leqslant s\leqslant m}2^{m-s}[C(m,s) - \sum_j C(|PP_c(S)[j]|,s)] \ .$$

Using the same identity as above, we can write this as

$$G_{|P|}(\{N - lub_i(S)\})(3^m - 2^m - m2^{m-1} - \sum_j 2^{m-q}(3^q - 2^q - q2^{q-1}))$$

(where j ranges over members of the partition $PP_c(S)$ whose size $q = |PP_c(S)[j]| > 1$), and s_n is the sum of these contributions for all non-root clusters of T_c.

To verify the correctness of this method, recognize that the maximal twigs less than N are exactly the elements of $PP_c(N)$. In one step of the algorithm, we compute the number of twigs (and their associated nests) that join at one (S) of these maximal twigs Then, to count the number of twigs restricted to just one child of S, we descend and locate the maximal twigs less than S: the elements of $PP_c(S)$. Thus the operation of the consensus algorithm produces exactly the clusters needed. ●

References

ADAMS, E N III (1972), "Consensus Techniques and the Comparison of Taxonomic Trees," *Systematic Zoology, 21,* 390-397

DAY, W H E (1983), "The Role of Complexity in Comparing Classifications," *Mathematical Biosciences, 66,* 97-114

DAY, W H E (1985), "Optimal Algorithms for Comparing Trees with Labeled Leaves," *Journal of Classification, 2,* 7-28

DAY, W H E, and MCMORRIS, F R (1985), "A Formalization of Consensus Index Methods," *Bulletin of Mathematical Biology, 47,* 215-229

ESTABROOK, G F , MCMORRIS, F R , and MEACHAM, C A (1985), "Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units," *Systematic Zoology, 34,* 193-200

MARGUSH, T , and MCMORRIS, F R (1981), "Consensus N-Trees," *Bulletin of Mathematical Biology, 43,* 239-244

NELSON, G (1979), "Cladistic Analysis and Synthesis: Principles and Definitions, with a Historical Note on Adanson's *Familles des Plantes* (1763-1764)," *Systematic Zoology, 28,* 1-21

NEUMANN, D A (1983), "Faithful Consensus Methods for n-Trees," *Mathematical Biosciences, 63,* 271-287

ROBINSON, D F (1971), "Comparison of Labeled Trees with Valency Three," *Journal of Combinatorial Theory, 11,* 105-119

SOKAL, R R , and ROHLF, F J (1981), "Taxonomic Congruence in the Leptopodomorpha Re-examined," *Systematic Zoology, 30,* 309-325

STINEBRICKNER, R (1984), "s-Consensus Trees and Indices," *Bulletin of Mathematical Biology, 46,* 923-935