Eurographics Conference on Visualization (EuroVis) 2013
B. Preim, P. Rheingans, and H. Theisel
(Guest Editors)

Volume 32 (2013), Number 3

Maximum Entropy Summary Trees

Howard Karloff[1] and Kenneth E. Shirley[1]

[1] AT&T Labs Research, Florham Park, NJ, USA

Abstract

Given a very large, node-weighted, rooted tree on, say, n nodes, if one has only enough space to display a k-node summary of the tree, what is the most informative way to draw the tree? We define a type of weighted tree that we call a summary tree *of the original tree that results from aggregating nodes of the original tree subject to certain constraints. We suggest that the best choice of which summary tree to use (among those with a fixed number of nodes) is the one that maximizes the information-theoretic entropy of a natural probability distribution associated with the summary tree, and we provide a (pseudopolynomial-time) dynamic-programming algorithm to compute this maximum entropy summary tree, when the weights are integral. The result is an automated way to summarize large trees and retain as much information about them as possible, while using (and displaying) only a fraction of the original node set. We illustrate the computation and use of maximum entropy summary trees on five real data sets whose weighted tree representations vary widely in structure. We also provide an additive approximation algorithm and a greedy heuristic that are faster than the optimal algorithm, and generalize to trees with real-valued weights.*

Categories and Subject Descriptors (according to ACM CCS): I.2.8 [Problem Solving, Control Methods, and Search]: Dynamic Programming—G.2.2 [Graph Theory]: Trees—I.4.10 [Image Representation]: Hierarchical—G.2.1 [Combinatorics]: Combinatorial Algorithms—

1. Introduction

Many data sets can be represented by a rooted, node-weighted tree, including employee organizational charts, web traffic logs, hard disk file structures, and phylogenetic trees, for example. The node weights can correspond to some node attribute of interest, or, in the absence of attributes, all the node weights can be set to one. Modern data sets represented by such trees can contain hundreds of thousands, or even millions, of nodes, so that visualizing them is challenging, and has received a great deal of interest in the research community (see [vLKS*11] for a thorough recent survey).

A natural goal of visualizing node-weighted trees is to be able to compare node weights across different nodes and branches of the tree while preserving a sense of the hierarchy, or structure, of the tree. Treemaps [Shn92] succeed in making comparisons of node weights easy, and they have been used for trees with as many as a million nodes [FP02], but they generally do a poor job of representing the visual hierarchy of the tree. On the other hand, traditional layered layouts succeed in displaying the tree's hierarchy, but require some additional visual encoding of node attributes (such as color, shape, or size) to allow for attribute comparisons. Most importantly, they are not scalable, and typically become impossible to fit onto a single page or screen if the tree of interest has more than a few hundred nodes.

In this paper we propose a method for visualizing large, node-weighted (unordered) rooted trees that allows comparisons of node attributes *and* preserves the visual hierarchy of the tree. We do this in three steps:

1. *Aggregation:* First, we define a novel way to aggregate nodes of a node-weighted tree that results in a new, smaller node-weighted tree that we call a *summary tree* of the original tree, whose number of nodes can be chosen to be any integer between one and the number of nodes in the original tree.
2. *Optimization:* Second, we provide an algorithm to compute the optimal summary tree out of all possible summary trees with a given number of nodes, where optimality is defined in terms of maximizing the information-

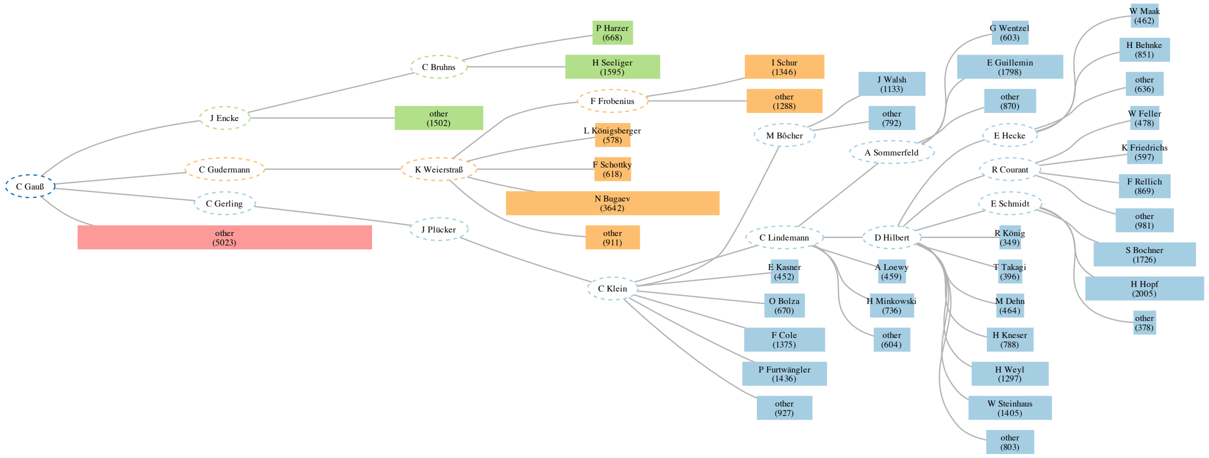

Figure 1: *The maximum entropy 56-node summary tree of the math genealogy tree rooted at Carl Friedrich Gauss, which has 43,527 equal-weighted nodes (where the original advisor-student graph was forced to be a tree by choosing the primary advisor for each student who had multiple advisors). Node colors are determined by their depth-1 ancestor, and node areas are proportional to their weights in the summary tree. This tree is best viewed on a computer screen.*

theoretic entropy of a natural distribution associated with the tree.

3. *Layout:* Last, we recommend that the optimal summary tree with a given number of nodes be visualized using a layered layout, where the node sizes are drawn in proportion to their weights. This type of layout is not required, but we feel it maximizes the utility of our methodology.

The resulting visualization is a *maximum entropy summary tree* of any order (i.e., number of nodes) between one and a user-specified bound $K \leq n$ that automatically provides the most informative summary of the original n-node tree among all summary trees of the chosen order. From our experiments on real-world data, we find that we can often compute a summary tree that is nearly as informative as the original tree (in terms of entropy) and which contains only a small fraction of the number of nodes of the original tree (often less than 100), thus easily fitting onto a single screen or page using a layered layout. In other words, our algorithm to compute maximum entropy summary trees is essentially a data reduction method that yields good visualizations, from both an aesthetic point of view and an information-theoretic point of view. See Figure 1 for an example applied to a mathematical genealogy tree [Mat], which is discussed in more detail later.

2. Background and contributions

The recent survey on techniques for drawing large graphs by von Landesberger et al. [vLKS*11] notes that the two basic types of tree visualization methods are space-filling layouts and node-link layouts. It is well-known that treemaps [Shn92], which are space-filling layouts, allow users to visually compare attribute values across nodes, and are scalable to trees with at least approximately one million nodes [FP02]. In the absence of attribute values, treemaps still allow users to compare the sizes of different branches of a tree by setting all node weights to one. The main weakness of treemaps, though, as pointed out by multiple authors, is that they do a poor job of showing the hierarchy, or structure, of a tree [vWvdW99, HMM00, BMH05, ZMC05].

Node-link diagrams, on the other hand, typically do a better job of displaying a tree's hierarchy, but are not necessarily scalable to trees with hundreds or thousands of nodes. Layered layouts, in which nodes on the same level of a tree are drawn along parallel lines, are especially conducive to showing the hierarchy of a tree, and are, unfortunately, especially difficult to scale to large trees, because the number of nodes at a given depth increases exponentially with the depth of most real-world trees [ZMC05].

A solution to the lack of scalability of node-link diagrams that many researchers have built on is the Focus+Context paradigm, in which a user interactively selects a region of a visualization to focus on, and the rest of the visualization is transformed, but still pictured, to provide context to the focus region. Hyperbolic browsers [LRP95] apply this paradigm to trees using hyperbolic geometry and a circular layout. A layered approach is the *accordion drawing technique* [MGT*03, BMH05], which uses "stretch-and-squish" navigation to allow users to browse large trees.

Another way to interactively apply Focus+Context techniques to large trees is to aggregate nodes. SpaceTree [GPB02], and Degree-of-Interest trees [CN02, HC04] combine node aggregation with a layered layout, where a degree-of-interest function determines which nodes are displayed or aggregated based on how “interesting" they are relative to the focal node. TreeWiz [RBB02] and Expand-Ahead methods [MDB04] also use aggregation and interaction to summarize and visualize large trees. Finally, *elastic hierarchies* [ZMC05] use a Focus+Context approach to allow users to manipulate hybrid visualizations that combine node-link diagrams and treemaps.

Our position is that (1) data reduction is an important first step in visualizing large trees and (2) theoretical principles should dictate which nodes are chosen for display. This second principle differs from much previous work, in which users interactively influence which nodes are displayed.

Regarding data reduction, we follow Herman et al., who write [HMM00] “... beyond a certain limit, no algorithm will guarantee a proper layout of large graphs. There is simply not enough space on the screen. In fact, from a cognitive perspective, it does not even make sense to display a very large amount of data. Consequently, a first step in the visualization process is often to reduce the size of the graph to display. As a result, classical layout algorithms remain usable tools for visualization, but only when combined with these techniques." We view summary trees as precisely this kind of data reduction technique. Our first main contribution is the definition of a summary tree, which is a novel way to aggregate nodes so that a very large node-weighted tree can be summarized by a potentially much smaller node-weighted tree *of any order that the user chooses*. The freedom to choose the order of a summary tree is an important property, because it is analogous to flexible zooming, and is a consequence of the specific constraints we impose on the node aggregation process.

Our second main contribution is introducing the notion of entropy to node-weighted trees. We define the entropy of a node-weighted tree as the information-theoretic entropy of a discrete probability distribution whose probabilities are defined by the normalized node weights. This is a natural way to think about the information contained in a node-weighted tree. Given a constraint on the number of nodes to display in a summary tree, we propose that the optimal choice of which fixed-order summary tree to display, among many possible choices, is the one with maximum entropy, because it is theoretically the most informative. We provide an exact algorithm to compute this summary tree for trees with nonnegative integral weights, and an approximation algorithm and a heuristic for the more general case of trees with nonnegative real weights. We recommend (but our algorithm does not require) that the nodes of a maximum entropy summary tree be drawn in a layered node-link diagram (preserving the visual hierarchy), with their sizes proportional to their weights (as in the case of treemaps, allowing for visual comparisons of attributes and tree substructure).

3. Summary trees

Given a rooted, node-weighted tree T with n nodes, we introduce the concept of a “summary tree" T' of T, which is a rooted, node-weighted tree with k nodes, where $1 \leq k \leq n$. Denote the weight of node i by w_i, where w_i is a nonnegative real number. One property of a summary tree is that each node of the summary tree T' is a nonempty subset of the node set $V(T)$ of T, the collection of nodes in T' being a partition of $V(T)$. The weight of a subset of nodes is defined to be the sum of the weights of the nodes in the subset. Also, given a node v of T, let T_v denote the subtree of T rooted at v.

Definition 1. *Given a rooted, node-weighted n-node tree T, a k-node* summary tree *T' of T is a rooted, weighted, k-node tree in which each node is a subset of $V(T)$, defined recursively as follows:*

1. *If T has exactly one node, v, then the unique summary tree of T is a 1-node tree whose one node is $\{v\}$ (and hence k must equal 1).*
2. *Suppose T has root v and children $v_1, v_2, ..., v_d$, $d \geq 1$. Then a summary tree T' of T is either*
   a. one node $V(T)$, or
   b. i. a root $\{v\}$,
      ii. a subset S_v of v's children, where if $S_v \neq \emptyset$, T' contains one node labeled “other," which equals $\cup_{x \in S_v} V(T_x)$, and
      iii. separate summary trees for T_{v_i} for all $v_i \notin S_v$.

   In case 2.b., the root $\{v\}$ is the parent in T' of the node labeled “other," if it exists, and of the roots of the summary trees for the T_{v_i}'s for $v_i \notin S_v$.

It is easy to see that the collection of sets represented by all nodes of T' is a partition of $V(T)$. It follows that the total weight of a summary tree T' of T is the same as the total weight of T. Note that the sets S_v are part of the definition of the tree. Hence if T is a 3-node tree on $\{a,b,c\}$ rooted at a (i.e., with edges $\{a,b\}$ and $\{a,c\}$), then there are three distinct 3-node summary trees of T. All three are isomorphic, but one tree has $S_a = \emptyset$, one has $S_a = \{b\}$, and one has $S_a = \{c\}$. This situation is caused only by the existence of an S_u of size one for some u.

The intuition behind summary trees is that they allow nodes to be aggregated in two useful ways, described by parts 2.a. and 2.(b.)ii. of Definition 1. One way (part 2.a. of Definition 1) is for a node of a summary tree to represent a whole subtree of the original tree. This is a common method of node aggregation for trees used by others [RBB02, CN02, GPB02, HC04].

The other way to aggregate nodes (part 2.(b.)ii. of Definition 1) is slightly more subtle—an “other" node in a summary tree represents a set of siblings from the original tree and all the descendants of those siblings—but a parent in the summary tree can have at most one such node among its children. There are two important motivating principles behind this type of aggregation. First, when a node has many children whose weights have a skewed distribution, it can be very useful to view the children with large weights individually (and possibly some of their descendants, too), while aggregating all the remaining children and their descendants into one node called “other" to save space. Second, we choose to restrict a node from having two or more “other" children. We argue that if multiple “other" nodes were desired under a single parent, then the attribute that distinguishes them from each other should be encoded into the hierarchy of the tree, defining a new split along the branch. If no such attribute exists, then only one “other" node is required. This restriction may simply be a matter of taste, but we feel it is consistent with the node aggregation theories described in [EF10]. DOITrees [HC04] include the notion of an “other" node, but it is not formally defined, and to our knowledge, it has not been discussed elsewhere.

One useful consequence of allowing “other" nodes in summary trees is that doing so guarantees the existence of a k-node summary tree of an n-node tree for each $k = 1,2,...,n$, which can be easily proven by induction on k. This property provides an analogue to flexible zooming, in which a user can view a sequence of k-node summary trees from $k = 1$ to a user-determined $K \leq n$. We in fact recommend this procedure as a way to explore the structure of a large node-weighted tree.

Last, Figure 2 illustrates the definition of a summary tree by showing a 9-node weighted tree and two 6-node summary trees of it. These trees, and all other trees we visualize in this paper, were drawn using the DOT algorithm in Graphviz [GKNpV93]. We draw them with a layered layout, in which the nodes are rectangular, with constant height and width proportional to their weight, or vice versa, so that their areas are proportional to their weights.

4. The entropy of a tree

Here we formally define the entropy of a sequence and of a node-weighted tree, and we introduce an important equation and a new definition that are used in our algorithm for computing maximum entropy summary trees.

We define the entropy of a sequence of nonnegative reals:

$$H(w_1, w_2, ..., w_n) = -\sum_{i=1}^{n} \left(\frac{w_i}{W}\right) \log_2\left(\frac{w_i}{W}\right), \qquad (1)$$

where W denotes the sum of the reals, if $W > 0$, and 0 otherwise. We take $0\log_2(0)$ to be 0 in this computation. (Also, from here onward, we denote “$\log_2$” by “lg.”) We define the entropy of a node-weighted, n-node tree T with node weights $w_1, w_2, ..., w_n$, to be $H(w_1, w_2, ..., w_n)$. We will also use the shorthand notation $H(T)$ to denote $H(w_1, ..., w_n)$.

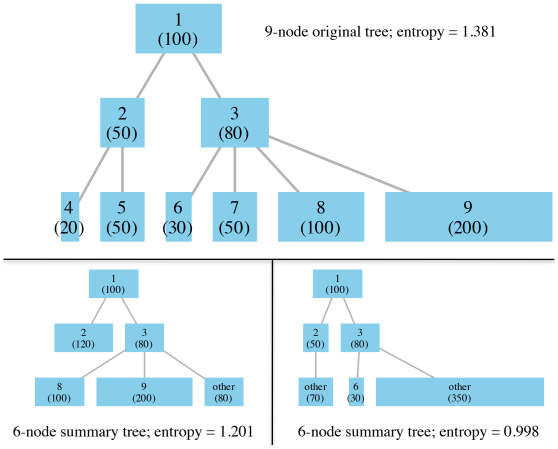

Figure 2: *In the upper panel, a 9-node tree (with node weights in parentheses), and below it, two different 6-node summary trees of the original 9-node tree, with their entropies (to be defined in Section 5) included.*

The justification for maximizing entropy is simple: given a fixed number of nodes to display, we wish to display the set of nodes that provides the most information about the distribution of node weights to the viewer. It would not be very informative, for example, to summarize a 10,000-node tree with a 50-node summary tree in which 99% of the weight of the tree is aggregated into one supernode, and the other 49 nodes only share 1% of the original tree's weight, if another more balanced aggregation were possible. Given such a “lopsided" summary, a user would naturally want to disaggregate the large supernode to learn its substructure, at the expense of aggregating some of the 49 small nodes. This intuition agrees with maximizing entropy, since entropy is maximized when all the weights are identical. Maximizing other objective functions besides entropy (e.g., $-\sum p_i^2$) is a potential direction for future work.

Next, we need to establish a fact about entropy. Suppose we have two discrete probability distributions (on nonoverlapping sets), with associated probabilities $p_1, ..., p_{k_1}$ and $p'_1, ..., p'_{k_2}$, and entropies h_1 and h_2. If we randomly choose an outcome from the first distribution with probability q or an outcome from the second distribution with probability $1-q$, then the resulting probability distribution (with $k_1 + k_2$ possible discrete outcomes) has entropy

$$h = qh_1 + (1-q)h_2 - q\lg(q) - (1-q)\lg(1-q). \qquad (2)$$

The important part of this result related to the dynamic-programming algorithm is that to compute entropy of the “combined" distribution, one does not need to know the specific probabilities associated with the two original distributions—one only needs the entropies of those two distributions, and the probability of choosing from one distribution versus the other. We use this result to define a function, combine$()$, that we will use in the algorithm.

Definition 2. *Where w_1, w_2 are nonnegative reals, not both zero, and $h_1, h_2 \geq 0$, let $q = w_1/(w_1 + w_2)$, and combine$(h_1, w_1; h_2, w_2) = qh_1 + (1-q)h_2 - q\lg(q) - (1-q)\lg(1-q)$. Define combine$(h_1, 0; h_2, 0) = 0$.*

Last, we introduce the idea of a "summary forest." Suppose a node v of T has children $v_1, v_2, ..., v_d$, where $d \geq 1$. Define a k-node *summary forest* for $T_{v_1} \cup T_{v_2} \cup \cdots \cup T_{v_l}$, for $1 \leq l \leq d$, as the forest that remains after removing $\{v\}$ from a $(k+1)$-node summary tree for the subtree of T consisting of v, its first l children, and their descendants. This collection of k nodes that we define as a summary forest is, in fact, a summary tree according to Definition 1 if $l = 1$, and is not a summary tree (because it's not connected) if $l > 1$. Further, if the $(k+1)$-node summary tree for the subtree of T consisting of v, its first l children, and their descendants contains an "other" child of v (i.e. $S_v \neq \emptyset$ from Definition 2.(b).ii.), then we still refer to this node as an "other" child of v in the summary forest. Last, if the weights of the nodes in the k-node summary forest are denoted $w_1, ..., w_k$, define the entropy of the summary forest as $H(w_1, ..., w_k)$.

5. An algorithm for finding maximum entropy summary trees

Given an n-node tree T and a target $K \leq n$, we introduce a pseudopolynomial-time dynamic-programming algorithm that computes the maximum entropy k-node summary trees for $k = 1, 2, ..., K$, provided that the node weights are integral. In Sections 6 and 8 we describe an approximation algorithm and a greedy heuristic that work for nonnegative real weights. The exact algorithm of this section runs in (truly) polynomial time when the sum W of the weights is small, e.g., when all node weights are 1, but not when W is large. (The fact that there are $2^d - 1$ possibilities for an "other" child of a parent with d children makes finding a polynomial-time algorithm difficult. Indeed, we leave existence of a truly polynomial-time algorithm for large W as an open problem.)

5.1. The recurrence

Our algorithm depends on one main idea. For a node v with d children $v_1, v_2, ..., v_d$, if we have the entropies of the k-node maximum entropy summary trees for the tree rooted at each child, for $k = 1, ..., K$, then we will compute the maximum entropy k-node summary tree for T_v for all k. To compute this via a recurrence, though, we must parameterize by the weight w of the "other" child of v.

Definition 3. *Let v be a node of T. We define $f_v(k, w)$ for $1 \leq k \leq K$, $-1 \leq w \leq W$ and $F_v(k)$ for $1 \leq k \leq K$.*

1. *For $w = 0, 1, 2, ..., W$, for $1 \leq k \leq K$, $f_v(k, w)$ is the maximum entropy of a k-node summary tree for T_v in which there is an "other" child of the node $\{v\}$ with weight w.*
2. *For $1 \leq k \leq K$, $f_v(k, -1)$ denotes the maximum entropy of a k-node summary tree for T_v in which there is no "other" child of the node $\{v\}$.*
3. *For $1 \leq k \leq K$, let $F_v(k) = \max_{w=-1}^{W} f_v(k, w)$, the maximum entropy of any k-node summary tree for T_v.*

Definition 4. *Fix $1 \leq l \leq d$, $1 \leq k \leq K-1$, and $-1 \leq w \leq W$. Let $g_v(l, k, w)$ be the maximum entropy of a k-node summary forest for $T_{v_1} \cup T_{v_2} \cup \cdots \cup T_{v_l}$ which contains an "other" child of v of weight w, if $w \geq 0$, or has no "other" child of v, if $w = -1$.*

To illustrate this definition, we compute $g_v(l, k, w)$ for the tree drawn in Figure 3 for $l = 4$, $k = 5$, and $w = 36$ (and $d = 6$). The only way to get an "other" child of v of weight 36 in the summary forest for $T_{v_1} \cup T_{v_2} \cup T_{v_3} \cup T_{v_4}$ is for the set S_v of children forming the "other" child to equal $\{1, 3\}$ or $\{2, 3\}$.

- If the "other" child consists of $S_v = \{1, 3\}$: We need $k = 5$ nodes from the proper descendants of v, one of which is the "other" child, so we need four non-"other" nodes. We can get four nodes from $V(T_{v_2})$ and $V(T_{v_4})$ by getting:
  - one from $V(T_{v_2})$ and three from $V(T_{v_4})$: entropy is $H(15+21, 3+5+7, 5, 10, 11) = 2.012198$; or
  - two from $V(T_{v_2})$ and two from $V(T_{v_4})$: entropy is $H(15+21, 3, 5+7, 5, 10+11) = 1.880552$; or
  - three from $V(T_{v_2})$ and one from $V(T_{v_4})$: entropy is $H(15+21, 3, 5, 7, 5+10+11) = 1.794777$.
- If the "other" child consists of $S_v = \{2, 3\}$: Again we need four non-"other" nodes. We can get four nodes from $V(T_{v_1})$ and $V(T_{v_4})$ by getting:
  - one from $V(T_{v_1})$ and three from $V(T_{v_4})$: entropy is $H(15+21, 2+6+7, 5, 10, 11) = 2.012198$;
  - two from $V(T_{v_1})$ and two from $V(T_{v_4})$: entropy is $H(15+21, 2, 6+7, 5, 10+11) = 1.850276$; or
  - three from $V(T_{v_1})$ and one from $V(T_{v_4})$: entropy is $H(15+21, 2, 6, 7, 5+10+11) = 1.77990$.

Hence $g_v(4, 5, 36)$ is the maximum of these six quantities and is equal to 2.012198, achieved in two ways.

Let the *size* s_v of v denote the sum of the weights of all the descendants of node v.

Lemma 1. *(Basis) Let v_1 denote the first child in an arbitrary ordering of v's children.*

1. *$g_v(1, 1, -1) = 0$ (i.e., there is a 1-node summary forest, having entropy 0 and having no "other" child of v, for the subtree rooted at the first child of v).*
2. *If $w \geq 0$, then $g_v(1, 1, w) = -\infty$, except that $g_v(1, 1, s_{v_1}) = 0$ (i.e., the only 1-node summary forest for the subtree rooted at v_1 which has an "other" child of v consists solely of an "other" child representing v_1 and all its descendants).*
3. *If $k > 1$, then $g_v(1, k, -1) = F_{v_1}(k)$ (i.e., the entropy of the maximum entropy summary forest for T_{v_1} with $k > 1$ nodes which has no "other" child of v has entropy $F_{v_1}(k)$, by definition of $F_{v_1}(k)$).*

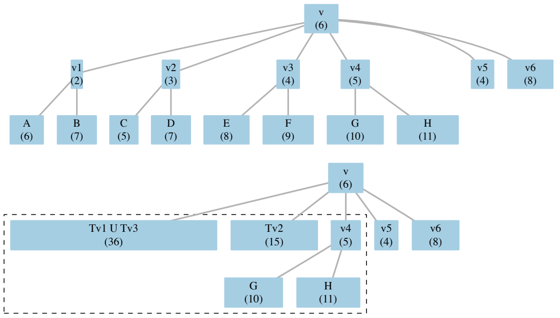

Figure 3: *The subtree, T_v, for which we illustrate the definition of $g_v(l,k,w)$, for $l=4$, $k=5$, $w=36$ (and $d=6$). Node weights are listed in parentheses. The upper figure is T_v, and the lower figure shows the maximum entropy summary forest for $T_{v_1} \cup \cdots \cup T_{v_4}$ for $k=5$ with an "other" child of v of size $w=36$ within the dashed box (where we chose the "other" child $S_v = \{1,3\}$).*

4. *If $k > 1$ and $w \geq 0$, then $g_v(1,k,w) = -\infty$ (i.e., any summary forest of T_{v_1} with two or more nodes cannot be part of an "other" child of v).*

Now we induct on l. The following lemma shows how to compute the $g_v(l,k,w)$ values from the $g_v(l-1,k,w)$ values. The inductive step applies this recurrence for $l = 2,3,4,\ldots,d$.

Lemma 2. *(Inductive Step)*

1. *For $l \geq 2$, $g_v(l,1,s_{v_1}+s_{v_2}+\cdots+s_{v_l}) = 0$, and $g_v(l,1,w) = -\infty$ for all $w \neq s_{v_1}+\cdots+s_{v_l}$. (Here, the first l children form an "other" child of v).*
2. *For $l \geq 2$, for all $k \geq 2$, $g_v(l,k,-1) = \max_{k_1=1}^{k-1} combine(g_v(l-1,k_1,-1), s_{v_1}+s_{v_2}+\cdots+s_{v_{l-1}}; F_{v_l}(k-k_1), s_{v_l})$. (There is no "other" child of v, and we combine a k_1-node summary forest for $T_{v_1} \cup \cdots \cup T_{v_{l-1}}$ containing no "other" child of v with a $(k-k_1)$-node summary tree for T_{v_l}).*
3. *For $l \geq 2$, for all $k \geq 2$, and $w \geq 0$, $g_v(l,k,w)$ is the maximum of the following three quantities:*
   a. *$\max_{k_1=1}^{k-1} combine(g_v(l-1,k_1,w), s_{v_1}+\cdots+s_{v_{l-1}}; F_{v_l}(k-k_1), s_{v_l})$, if $w \leq s_{v_1}+s_{v_2}+\cdots+s_{v_{l-1}}$, and $-\infty$ otherwise.*
   b. *$combine(g_v(l-1,k-1,-1), s_{v_1}+\cdots+s_{v_{l-1}}; 0, s_{v_l})$, if $s_{v_l} = w$ (and is $-\infty$ otherwise).*
   c. *$\frac{-1}{M+s_{v_l}}\left[(-MH + M\lg M - (w-s_{v_l})\lg(w-s_{v_l})) - (M+s_{v_l})\lg(M+s_{v_l}) + w\lg w\right]$, where $M = s_{v_1}+\cdots+s_{v_{l-1}}$, and $H = g_v(l-1,k,w-s_{v_l})$, if $w - s_{v_l} \geq 0$ (and $-\infty$ otherwise).*

For lack of space, the proof appears in the appendix. We mention here only that case 3(c) is interesting because we "merge" T_{v_l} into an existing "other" child of v in the summary forest for $T_{v_1} \cup \cdots \cup T_{v_{l-1}}$. The entropy calculation in equation (2) does not apply; hence there is a need for a new formula.

Recall that v has d children. When we finish with this induction on l, we have all $g_v(d,k,w)$ values. Given that the $g_v(d,k,w)$'s are defined for the summary forest for $T_{v_1} \cup \cdots \cup T_{v_d}$, the only node missing from the subtree rooted at v is v itself, so to get the $f_v(k,w)$'s, we simply have to "attach the root." This is easy. The proof is omitted.

Lemma 3. *("Attaching the root")*

1. *$f_v(1,-1) = 0$ and $f_v(1,w) = -\infty$ for all $w \geq 0$.*
2. *If $k \geq 2$, $-1 \leq w \leq s_{v_1}+s_{v_2}+\cdots+s_{v_d}$, then $f_v(k,w) = combine(0, w_v; g_v(d,k-1,w), s_{v_1}+\cdots+s_{v_d})$.* ■

5.2. The algorithm

Given the recurrence of the previous section, creating an algorithm for the case of nonnegative integral weights is easy. One can process the nodes in nonincreasing order by depth, computing $F_u(k)$ for all k for all children u of a node v before computing $F_v(k)$ for any k. To compute $F_v(k)$ for a node v and all k's, one computes $f_v(k,w)$ for all k and w. One does this by computing $g_v(l,k,w)$ for all k and w, for $l = 1,2,\ldots,d$ (where v has d children), in that order, via Lemma 1 for the basis and Lemma 2 for the recurrence.

Here is pseudocode for computing the optimal entropies. (How to generate the trees is easy and is omitted.)

```
• For $v \in \{1,\ldots,n\}$ in nonincreasing order by $depth(v)$, do:
    – If $v$ is a leaf, set $F_v(1) = 0$ and $F_v(k) = -\infty$ for $k = 2,3,\ldots,K$.
    – Else do
        ○ Where $v$ has $d \geq 1$ children, use Lemma 1 to define $g_v(1,k,w)$ for all $k,w$.
        ○ For $l = 2,3,\ldots,d$, do:
            ◇ Use Lemma 2 to compute $g_v(l,k,w)$ for all $k,w$.
        ○ (Attach $v$:) Use Lemma 3 to compute $f_v(k,w)$ for all $k,w$.
        ○ Set $F_v(k) = \max_{w=-1}^{W} f_v(k,w)$ for all $k$.
• Output $F_1(k)$ for all $k$.
```

The time needed by the algorithm is $O(K^2nW)$, which is pseudopolynomial in the input size. (A polynomial-time algorithm would run in time polynomial in n and $\lg W$, since W can be represented in binary in approximately $\lg W$ bits.) Unfortunately we do not know if the problem is NP-Complete.

To illustrate the result, Figure 4 displays the maximum entropy 60-node summary tree for a company organizational chart with 43,134 employees. The structure of the organization is clear: there are five main branches, where the blue- and green-colored branches are the largest. Some employees at depth 3 (such as employee 265, the second-rightmost blue node) have many more employees under them than employees at depth 2. The summary tree pictured has maximum

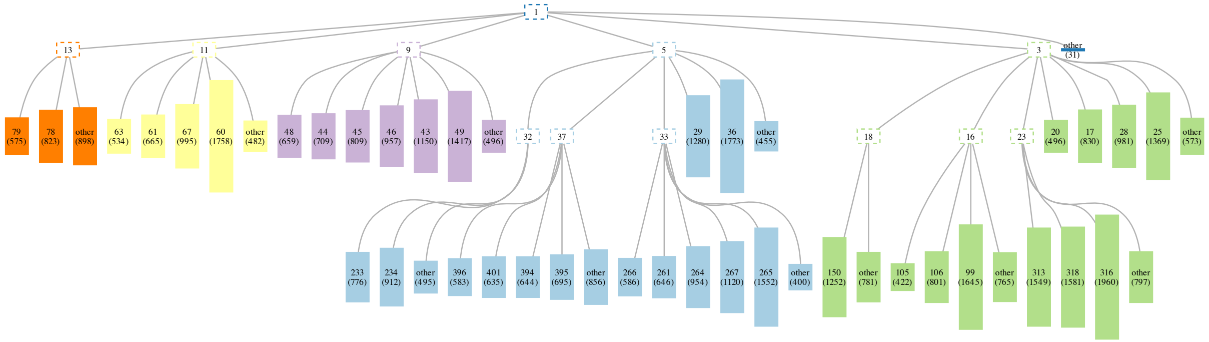

Figure 4: *The maximum entropy 60-node summary tree of a company organizational chart that has 43,134 equal-weighted nodes. Nodes are labeled 1 through* $n = 43,134$*, node colors are determined by their depth-1 ancestor, and node areas are proportional to their weights in the summary tree, which are labeled in parentheses, except summary tree nodes with weight 1, where the node is drawn transparently with a dotted outline.*

entropy, and therefore theoretically provides the viewer the most information about the distribution of node weights among all 60-node summary trees.

6. A polynomial-time additive approximation algorithm

What can one do if the weights are reals, which are forbidden by the dynamic-programming algorithm which computes an optimal summary tree? Even if the weights are all integral, what can one do if their sum W is huge? To address these two concerns, we give an additive approximation algorithm, which takes a tree T weighted with nonnegative real weights (w_i), a positive integer K, and an $\varepsilon > 0$, and produces, for each $k \leq K$, a k-node summary tree whose entropy is at most ε less than that of the optimal k-node summary tree.

However, for lack of space the full writeup of the algorithm appears in the appendix in Section 9. Here we only give a cursory summary.

The idea underlying the algorithm is simple: (1) scale the weights uniformly so that they sum to an integer W, whose value will be determined later; (2) carefully round real weight w_i to $w'_i \in \{\lfloor w_i \rfloor, 1 + \lfloor w_i \rfloor\}$; and (3) run the dynamic-programming algorithm of the previous section on the scaled, rounded weights. Doing so, however, in such a way as to guarantee small enough error relative to the maximum entropy of the original weights while simultaneously keeping the running time down is quite nontrivial. The rounding method uses elements of mathematical discrepancy theory [Spe94, Cha00]. Specifically, we round the w_i's to w'_i's such that for all nodes v in T, the sum of w_i over descendants i of v differs from the sum of w'_i over descendants i of v by at most 1 in absolute value. (Naively rounding each w_i up or down instead of minimizing the discrepancy on subtrees gives an algorithm approximately 1000 times slower on some of our data sets.) In addition, showing that such a rounding suffices to give entropy within ε of the optimal entropy for a suitable W (Lemma 5) is one of the more interesting results in this paper.

Here is the algorithm. Let us denote by T^w an n-node tree on $\{1, 2, ..., n\}$ whose ith node has real weight w_i.

1. Choose the least integer $W \geq 3$ such that $(2/\ln 2)(3K/W)(1 + \ln K - \ln(3K/W)) \leq \varepsilon$ and scale $\langle w_1, w_2, ..., w_n \rangle$ to have sum W.
2. Using our rounding algorithm (Lemma 4 in the appendix), produce a sequence $\langle w'_1, w'_2, ..., w'_n \rangle$ with $w'_i \in \{\lfloor w_i \rfloor, 1 + \lfloor w_i \rfloor\}$ such that for any node v in T, the sums of w_i over descendants i of v and of w'_i over descendants i of v differ by at most 1 in absolute value.
3. Run the exact dynamic-programming algorithm of Section 5.2 on tree $T^{w'}$, to get an optimal k-node summary tree T' for $T^{w'}$.
4. Output tree Z, which is T' except with weight w_i on node i instead.

It is not hard to see that the least W is $O((K/\varepsilon)\log(\max\{K, 1/\varepsilon\}))$ and independent of n.

Definition 5. *Let* $OPT_k(T^w)$ *denote the entropy of a maximum entropy* k*-node summary tree of* T^w.

Now we give the main theorem of this section.

Theorem 1. *The tree* Z *produced by the algorithm is a* k*-node summary tree for* T^w *having (binary) entropy at least* $OPT_k(T^w) - \varepsilon$*. The running time of the algorithm is* $O((K^3/\varepsilon)n\log(\max\{K, 1/\varepsilon\}))$.

Please see Section 9 in the appendix for details.

7. Examples

We illustrate the computation and visualization of maximum entropy summary trees on five real-world data sets that can be represented by large, rooted, node-weighted trees.

First, we consider a set of aggregated webpage visits to a large Internet portal by a sample of a million users on one day in March, 2012. The nodes of this tree are webpages that are organized hierarchically into categories (such as "/home/news/international/russia," for example, or "/home/sports/baseball"), and the weights are the number of clicks per webpage aggregated across all users. There are 19,335 nodes in this tree, with a depth of 17 levels, a range of zero to 365 children per node, and total weight of over 260 million. The distribution of weights per node is highly skewed, with one webpage receiving over 20% of all clicks, and a long tail in which 45% of webpages received 3 or fewer clicks.

Since the sum of the weights was very large (260 million), the exact algorithm of Section 5 was infeasible. Instead we use the approximate algorithm of Section 6, with $\varepsilon = 0.05$, to compute summary trees that are nearly optimal. We computed nearly optimal k-node summary trees for $k = 1,\dots,100$, and we view them in sequence to learn about the distribution of clicks across the taxonomy of the web portal. Figure 5 compares a summary tree computed using a naive aggregation of weights to the maximum entropy summary tree of the same order, and to a larger maximum entropy summary tree. The maximum entropy summary trees naturally aggregate nodes in a way that spreads out their weights as evenly as possible, resulting in informative visualizations.

The other data sets we investigated are:

1. One co-author's hard drive, which contains 15,671 files and directories, with node weights set to file sizes in kilobytes. This drive has a total of 143,990,819 kilobytes of disk space, where the tree has a depth of 6 levels, and the number of children per node ranges from zero to 5,342.
2. The phylogenetic tree data from the Tree of Life Web Project [MS07]. This tree has 94,080 nodes, 54,121 of which represent a species or subspecies (and were given a weight of 1), and the other 39,959 of which represent a taxonomic categorization (such as "animal" or "plant," and thus were given a weight of zero).
3. The Mathematics Genealogy Project [Mat] subtree rooted at Carl Friedrich Gauss, which has 43,527 nodes, all given a weight of 1. For students with multiple advisors, we forced the graph to be a tree by assigning the primary advisor as the parent.
4. A section of an employee organizational chart, from a large company, which contains 43,134 employees, all given a weight of one.

For all five data sets (these four plus the web traffic data), we computed four sets of k-node summary trees (for $k = 1,\dots,100$): (1) maximum entropy summary trees (when feasible), (2) and (3), approximately maximum entropy summary trees using $\varepsilon = 0.05$ and $\varepsilon = 0.1$, respectively, and (4) a greedy heuristic (which we describe in Section 8). Table 1 contains the running times for all four procedures for each

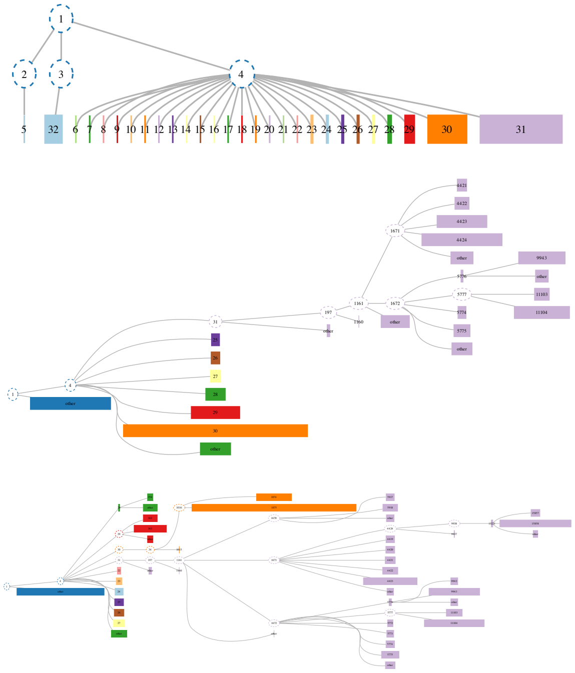

Figure 5: *Three summary trees of the 19,335-node web traffic tree. The upper figure is a naive aggregation to depth 2, where the node weights are heavily skewed. The middle figure is the maximum entropy 32-node summary tree, which displays much more information given the same number of nodes. The bottom figure is the maximum entropy 60-node summary tree, which provides an even finer-grained view of the structure of clicks across the taxonomy of the web portal. We color the nodes according to their depth-2 ancestor, and we draw their sizes proportional to their weights.*

data set, except the optimal algorithm for the web traffic and hard drive data sets, whose weights were too large for this algorithm to be feasible. Running times were longer for the approximation algorithm than the optimal algorithm for three data sets because the sum of the scaled weights, which depends only on K and ε, was higher than the sum of the *original* weights. In these cases, running the optimal algorithm is obviously preferable. We strongly encourage the reader to view the visualizations of these sets of summary trees in the supplementary materials, or on the author's website [Shi]. For each data set, it is very instructive to view the summary trees in sequence on a computer screen, from $k = 1,..,100$, to see the structure of the tree in increasing detail.

Another way to view the effectiveness of maximum entropy summary trees is to plot their entropies for successive values of k and compare them to the entropy of the original tree. Figure 6 illustrates this curve for each of the five examples.

Table 1: *Running times (on a 2.67 GHz machine with 48 GB of memory) in minutes and seconds for different algorithms on five real-world data sets, where columns labeled "*$\varepsilon =$*" refer to the approximation algorithm with the given values of* ε*, and* $K = 100$ *for each run. A hyphen indicates an instance which did not terminate.*

Data Set	n	W	Opt.	$\varepsilon = 0.05$	$\varepsilon = 0.1$	Greedy
Web Traffic	19335	260M	—	4:44	2:17	0:01
Hard Drive	15671	143M	—	3:49	2:13	0:01
Tree of Life	94080	54121	8:31	33:38	16:16	0:06
Math Gen.	43527	43527	2:23	10:47	5:24	0:02
Org Chart	43134	43134	2:39	11:45	6:12	0:03

We call this curve the *entropy profile* for a given weighted, rooted tree. In the case of the web traffic data (Figure 6(a), black line), the entropy profile shows that we can draw a summary tree with about 92.2% as much entropy as the original 19,335-node tree using only 100 nodes, which represents a great reduction in the size of the display for a small cost in terms of information loss. For the hard drive data, we achieve 91.1% of the entropy of the 15,671-node original tree with only 100 nodes. The other three data sets have much higher entropy in their original form, since all their weights are one (or zero, in the case of some of the Tree of Life nodes), and they naturally have high entropies. In these cases, it is instructive to compare the entropy profiles to the entropy of a uniform distribution with k categories, for $k = 1,\ldots,100$. This curve is illustrated by the green line in Figure 6(a) and (b). Even though the k-node maximum entropy summary trees for $k \le 100$ aren't obtaining a high fraction of the entropy of the original tree, they are—especially in the case of the organizational chart—achieving nearly as high an entropy (namely, $\lg k$) as possible for a discrete distribution with $k \le 100$ possible outcomes.

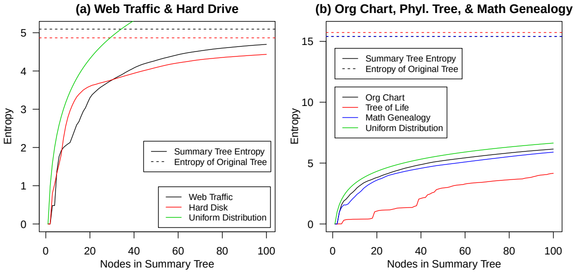

Figure 6: *Entropy profiles for* $k = 1,\ldots,100$ *for all 5 data sets. For the web traffic and hard drive data sets, the (approximate) maximum entropy summary trees have nearly as high entropy as the original trees using many fewer nodes.*

8. A (faster) heuristic

In this section we give a fast greedy heuristic allowing one to compute K-node summary trees in time $O(K^2 n)$, independently of W. While these summary trees are not optimal, in all real data sets we have tested, the heuristic always returned at least 94% of the optimal entropy, and never took more than six seconds.

Our greedy heuristic is motivated by a fact about depth-1 trees: if a maximum entropy summary tree of a depth-1 tree has an "other" node with, say, m leaves in it, then this "other" node must be comprised of the m leaves with least weight. Replacing one of the smallest m leaves in the "other" node with a larger leaf always yields a lower-entropy summary tree. This means for a depth-1 tree where the root has d children, the maximum entropy k-node summary tree will always have an "other" node comprised of the $d-k+2$ smallest leaves, for $2 \le k \le d$.

Our greedy heuristic extends this idea to the whole tree by considering only those "other" nodes comprised of the least-size children among siblings. (If a different set of "other" nodes uniquely produces the optimal summary tree, the greedy heuristic will return a suboptimal summary tree.) The greedy heuristic processes the nodes of a tree so that a node is processed after all its children and so that the children of a node are processed in nondecreasing order by size. We maintain an *entropylist* or *elist* for each node v. The *elist* for v is a sequence of K reals, the kth being the entropy of some k-node summary tree (ideally an optimal one, but not necessarily) of T_v (or $-\infty$ to correspond to no summary tree).

To explain the greedy algorithm, first we define a function combine_lists$(vec, \alpha; vec', \alpha')$, which takes two K-vectors vec, vec' and their respective weights $\alpha, \alpha' \ge 0$ and returns one K-vector vec_out:

1. $vec\_out_1 = 0$.
2. For $k = 2$ to K, $vec\_out_k = \max_{k_1=1}^{k-1} \text{combine}(vec_{k_1}, \alpha; vec'_{k-k_1}, \alpha')$.

Let z be a K-dimensional vector which is all $-\infty$'s, except with $z_1 = 0$. To process v:

1. If v is a leaf, set $elist_v = z$ and return.
2. Let v have children $v_1, v_2, \ldots, v_d$, sorted into nondecreasing order by size.
3. Generate a sequence $\langle L_v^1, L_v^2, \ldots, L_v^d \rangle$ of vectors in $\mathbb{R}^K$, as follows: $L_v^1 = elist_{v_1}$, and for $l = 2,3,\ldots,d$, $L_v^l = \text{combine\_lists}(L_v^{l-1}, s_{v_1} + s_{v_2} + \cdots + s_{v_{l-1}}; elist_{v_l}, s_{v_l})$.
4. (Now attach v:) $elist_v = \text{combine\_lists}(L_v^d, s_{v_1} + s_{v_2} + \cdots + s_{v_d}; z, w_v)$.

It is not hard to see that $elist_v[k]$, if nonnegative, is the entropy of a k-node summary tree for T_v. It is easy to prove that for binary trees, the greedy algorithm is optimal, since it can never miss an "other" node. The reader can find an instance for which the greedy algorithm generates a 4-node summary tree with only 2/3 of the optimal entropy in Section 9.3 in the appendix. An interesting open question is whether there is a $c > 0$ such that the entropy returned by the greedy algorithm is always at least c times optimal.

References

[BMH05] Beermann D., Munzner T., Humphreys G.: Scalable, robust visualization of very large trees. *Proc. EuroVis* (2005), 37–44. 2

[Cha00] Chazelle B.: *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, New York, NY, USA, 2000. 7, 11

[CN02] Card S. K., Nation D.: Degree-of-interest trees: a component of an attention-reactive user interface. *Proceedings of the Working Conference on Advanced Visual Interfaces* (2002), 231–245. 2, 3

[Doe04] Doerr B.: Linear discrepancy of totally unimodular matrices. *Combinatorica 24*, 1 (January 2004), 117–125. 12

[EF10] Elmqvist N., Fekete J.: Hierarchical aggregation for information visualization: Overview, techniques and design guidelines. *IEEE Transactions on Visualization and Computer Graphics 16*, 3 (2010), 439–454. 4

[FP02] Fekete J.-D., Plaisant C.: Interactive information visualization of a million items. *Proceedings of IEEE Symposium on Information Visualization* (2002), 117–124. 1, 2

[GKNpV93] Gansner E., Koutsofios E., North S. C., phong Vo K.: A technique for drawing directed graphs. *IEEE Transactions on Software Engineering 19* (1993), 214–230. 4

[GPB02] Grosjean J., Plaisant C., Bederson B.: Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. *Procedings of IEEE Symposium on Information Visualization* (2002), 57–64. 2, 3

[HC04] Heer J., Card S. K.: Doitrees revisited: Scalable, space-constrained visualization of hierarchical data. In *Advanced Visual Interfaces* (2004), pp. 421–424. URL: http://vis.stanford.edu/papers/doitrees-revisited. 2, 3, 4

[HMM00] Herman I., Melancon G., Marshall M. S.: Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics 6*, 1 (2000), 24–43. 2, 3

[LRP95] Lamping J., Rao R., Pirolli P.: A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1995), CHI '95, ACM Press/Addison-Wesley Publishing Co., pp. 401–408. URL: http://dx.doi.org/10.1145/223904.223956, doi:10.1145/223904.223956. 2

[Mat] The mathematics genealogy project [online]. URL: http://www.genealogy.ams.org/ [cited June 7, 2012]. 2, 8

[MDB04] McGuffin M. J., Davison G., Balakrishnan R.: Expand ahead: a space-filling strategy for browsing trees. *Proceedings of IEEE Symposium on Information Visualization* (2004), 119–126. 3

[MGT*03] Munzner T., Guimbretiere R., Tasiran S., Zhang L., Zhou Y.: Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *ACM Transactions on Graphics 22*, 3 (2003), 453–462. 2

[MS07] Maddison D. R., Schulz K. S.: The tree of life web project, 2007. URL: http://tolweb.org [cited May 3, 2012]. 8

[Nau04] Naudts J.: Continuity of a class of entropies and relative entropies. *Reviews in Mathematical Physics 16*, 6 (2004), 809–822. 11, 12, 13

[RBB02] Rost U., Bornberg-Bauer E.: Treewiz: interactive exploration of huge trees. *Bioinformatics 18*, 1 (2002), 109–114. 3

[Shi] Shirley K. E.: [online]. URL: www.research.att.com/~kshirley/summarytrees [cited April 10, 2013]. 8

[Shn92] Shneiderman B.: Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics 11*, 1 (1992), 92–99. 1, 2

[Spe94] Spencer J.: *Ten Lectures on the Probabilistic Method*, 2 ed. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1994. 7, 11

[vLKS*11] von Landesberger T., Kuijper A., Schreck T., Kohlhammer J., van Wijk J. J., Fekete J.-D., Fellner D. W.: Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum 30*, 6 (2011), 1719–1749. 1, 2

[vWvdW99] van Wijk J. J., van de Wetering H.: Cushion treemaps. *Proceedings of IEEE Symposium on Information Visualization* (1999), 73–78. 2

[ZMC05] Zhao S., McGuffin M. J., Chignell M. H.: Elastic hierarchies: Combining treemaps and node-link diagrams. *Proceedings of IEEE Symposium on Information Visualization* (2005), 57–64. 2, 3

9. Appendix: Supplementary Material

9.1. A polynomial-time additive approximation algorithm

9.1.1. High-level description

What can one do if the weights are reals, which are forbidden by the dynamic-programming algorithm which computes an optimal summary tree? Even if the weights are all integral, what can one do if their sum W is huge? To address these two concerns, in this section we give an additive approximation algorithm, which takes a tree T weighted with nonnegative real weights (w_i), a positive integer K, and an $\varepsilon > 0$, and produces, for each $k \leq K$, a k-node summary tree whose entropy is at most ε less than that of the optimal k-node summary tree.

The idea underlying the algorithm is simple—(1) scale the weights uniformly so that they sum to an integer W, whose value will be determined later; (2) carefully round real weight w_i to $w_i' \in \{\lfloor w_i \rfloor, 1 + \lfloor w_i \rfloor\}$; and (3) run the dynamic-programming algorithm of the previous section on the results—but doing so in such a way as to guarantee small enough error relative to the optimum for $\langle w_1, w_2, \ldots, w_n \rangle$ while simultaneously keeping the running time down is quite nontrivial. Larger W gives better accuracy at the cost of a larger running time. (We will always ensure that $\sum w_i' = \sum w_i = W$ to keep the normalization simple.) Specifically, we have to address two questions: (1) how does the entropy of a maximum entropy summary tree change if weights are rounded, and (2) how should one round the weights?

Our two-part answer to question (1) is given in Lemmas 5 and 6. Any k-node summary tree corresponds to a partition of the vertex set $V = V(T)$ into k parts. For a single fixed but *unknown* partition $\langle S_1, S_2, \ldots, S_k \rangle$, we are interested in two probability distributions. One, with W_j denoting $\sum_{i \in S_j} w_i$, is the probability distribution $\langle W_1/W, W_2/W, \ldots, W_k/W \rangle$. The second is the same with w_i' in place of w_i; specifically, it is $\langle W_1'/W, W_2'/W, \ldots, W_k'/W \rangle / W$, with $W_j' = \sum_{i \in S_j} w_i'$. We are interested in how much the entropies of the distributions $\langle W_1', W_2', \ldots, W_k' \rangle / W$ and $\langle W_1', W_2', \ldots, W_k' \rangle / W$ differ. Fortunately Lemma 6 [Nau04] bounds the entropy difference in terms of the L_1 distance between the distributions. Therefore, we ask, how can we round the weights to keep the L_1 distance $(\sum_{j=1}^{k} |W_j - W_j'|)/W$ small, *without knowing the partition* $\langle S_1, S_2, \ldots, S_k \rangle$ *in advance*?

The surprising result is that if one ensures that for any node v in the known input tree T, the sums of w_i/W and w_i'/W over descendants of v are almost the same, then for any *unknown* partition $\langle S_1, S_2, \ldots, S_k \rangle$ derivable from a k-node summary tree, the two induced probability distributions will have small L_1 distance $\sum_{i=1}^{k} |(W_i' - W_i)/W|$. This fact is proven in Lemma 5. The beauty here is that T is known in advance, whereas $\langle S_1, S_2, \ldots, S_k \rangle$ is not. We define *subtree absolute discrepancy* M to be the maximum, over nodes v, of the absolute difference between the sums of w_i and w_i' over descendants i of v.

This argument motivates the answer to (2): we should round the weights so that the subtree absolute discrepancy is small. How to do so is an interesting question in itself. There is much work on discrepancy theory for general set systems [Spe94, Cha00]. The surprising fact, known before but rediscovered together with an associated algorithm by the authors, is that one can round the w_i's while giving subtree discrepancy M bounded by 1, for any T.

Having said all that, the real argument is more complicated. There is no *single* partition $\langle S_1, S_2, \ldots, S_k \rangle$ with which one works. One has to argue that rounding weights from w_i to w_i' gives a new solution which is neither too large nor too small. To do this properly, one has to start the argument from the optimal partitions for both weights (w_i) and weights (w_i'). This argument is given in Lemma 7.

9.1.2. Details

Recall that we denote by T^w an n-node tree on $\{1, 2, \ldots, n\}$ whose ith node has real weight w_i.

We give an algorithm which takes a tree T^w on $\{1, 2, \ldots, n\}$, whose node i has nonnegative *real* weight w_i, positive integer K, and a positive real ε, and returns, for each $k \leq K$, a k-node summary tree whose entropy is at least $OPT_k(T^w) - \varepsilon$, where $\varepsilon > 0$ is a parameter. The running time of the algorithm (to generate all K trees) is $O((K^3/\varepsilon) n \log(\max\{K, 1/\varepsilon\}))$, though this is just a tree-independent worst-case upper bound.

In a rooted tree T, $x \in V(T)$, let T_x denote the subtree of T rooted at x.

Definition 6. *Suppose* (w_i), (w_i') *are both real-valued weight functions defined on* $\{1, 2, \ldots, n\}$.

1. *The* (signed) discrepancy $disc(S)$ *of a set* $S \subseteq \{1, 2, \ldots, n\}$ *is* $disc(S) = \sum_{i \in S} (w_i' - w_i)$.
2. *The* absolute discrepancy *of a set* $S \subseteq \{1, 2, \ldots, n\}$ *is* $|disc(S)|$.
3. *Relative to tree* T, *the* subtree absolute discrepancy M *is* $\max_i |disc(V(T_i))|$.
4. *Given an ordered partition* $P = \langle S_1, S_2, \ldots, S_k \rangle$ *of* $\{1, 2, \ldots, n\}$, *the* absolute discrepancy *of* P *is* $\sum_{i=1}^{k} |disc(S_i)|$.

Definition 7. *Say the pair* (w, w') *of weight functions is* nearby *if* $|w_i' - w_i| \leq 1$ *for all i.*

We start with our discrepancy lemma.

Lemma 4. *There is a* $O(n)$*-time algorithm that takes n and an n-node rooted tree T on* $\{1, 2, \ldots, n\}$ *rooted at node 1, and a sequence* $\langle w_1, w_2, \ldots., w_n \rangle$ *of nonnegative reals, and produces a sequence* $w_1', w_2', \ldots, w_n'$ *with* $w_i' \in \{\lfloor w_i \rfloor, 1 + \lfloor w_i \rfloor\}$ *such that the subtree absolute discrepancy M is at most 1. Furthermore, if the* w_i*'s sum to an integer, the* w_i'*'s will have the same sum.*

The existence of a rounding with subtree absolute discrepancy strictly less than 1 follows from a much more general result [Doe04], which itself follows on similar earlier results. The existence of the algorithm also probably follows from earlier results and will not be included here for lack of space.

Here is the algorithm.

1. Choose an integer W, as described later, and scale $\langle w_1, w_2, ..., w_n\rangle$ to have sum W.
2. Using Lemma 4, produce a sequence $\langle w'_1, w'_2, ..., w'_n\rangle$ with $w'_i \in \{\lfloor w_i \rfloor, 1+\lfloor w_i \rfloor\}$ having subtree absolute discrepancy $M \leq 1$ and with $\sum w'_i = W$.
3. Run the exact dynamic-programming algorithm of Section 5.2 on tree $T_{w'}$, to get an optimal k-node summary tree T' for $T_{w'}$.
4. Output tree Z, which is T' except with weights (w_i) instead. (In other words, output the same summary tree, but with the weight of a cluster containing nodes $S \subseteq \{1,2,...,n\}$ being $\sum_{i \in S} w_i$, instead of $\sum_{i \in S} w'_i$.)

Definition 8. *Say a node v in a summary tree T' is a* singleton *node if its cluster has size 1, is a* tree node *if its cluster has size exceeding 1 and it represents $V(T_x)$ for some node x, and otherwise is an "other" node.*

Note that any "other" cluster which corresponds to the descendants of exactly one child of a node v is being renamed a tree node or a singleton node for the purpose of this definition. Also note that every node in a summary tree is exactly one of singleton, tree, and "other."

Definition 9. *Say a node in a summary tree T' is* active *if it is not an "other" node. Let A_v be the set of active children of v in the summary tree T' and let $a_v = |A_v|$.*

It is obvious that $\sum_{v \in V(T')} a(v) \leq k$ if T' is a k-node summary tree, since $A_u \cap A_v = \emptyset$ for $u \neq v$ implies that $\sum_v |A_v| \leq k$.

Now we show that keeping small the subtree absolute discrepancy relative to T ensures that the absolute discrepancy of the partition associated with every k-node summary tree of T will be small.

Lemma 5. *Let T be a rooted tree on $V = \{1,2,...,n\}$ and let (w, w') be a nearby pair of weight functions on V. Let M be the subtree absolute discrepancy (relative to tree T) of that pair. Let $D = k + 2kM$. Let $P = \langle S_1, S_2, ..., S_k\rangle$ be the partition of V defined by any k-node summary tree T' for T. Then the absolute discrepancy of P is at most D.*

It is important for this lemma that P be derived from a k-node summary tree for T (and not be an arbitrary partition into k parts).

Proof. We need to prove that $\sum_{i=1}^{k} |disc(S_i)| \leq k + 2kM$, where $M = \max_{v \in V(T)} |disc(V(T_v))|$. Each set S_i corresponds to either a singleton node in T', a tree node in T', or an "other" node in T'.

If S_i corresponds to a singleton node in T', then $|S_i| = 1$ and, say, $S_i = \{u\}$. Then $|disc(S_i)| = |w_u - w'_u| \leq 1$, because (w, w') is nearby.

If S_i corresponds to a tree node in T', then there is a node $x \in V(T)$ such that $S_i = V(T_x)$ and $|disc(S_i)| = |\sum_{y \in V(T_x)} (w'_y - w_y)| \leq M$.

Now if u is an "other" node in T', which is cluster C in T, whose parent in T' is v, then $disc(V(T_v)) = (w'_v - w_v) + \sum_{a \in A_v} disc(V(T_a)) + disc(C)$, and therefore $disc(C) = disc(V(T_v)) - (w'_v - w_v) - \sum_{a \in A_v} disc(V(T_a))$. Hence $|disc(C)| \leq |disc(V(T_v))| + 1 + \sum_{a \in A_v} |disc(V(T_a))|$ $\leq M + 1 + a_v M$. Clearly this can be bounded by $1 + (k+1)M$, proving that $\sum_i |disc(S_i)| \leq k(1 + (k+1)M)$, but we can do better.

Let k_s be the number of singleton nodes in T', let k_t be the number of tree nodes in T', and let k_o be the number of "other" nodes in T'. Clearly $k_s + k_t + k_o = k$.

Any "other" cluster S has a parent node u in the summary tree. Let $parent(S)$ denote the parent of S, which is a singleton node. Hence $a_{parent(S)}$ denotes the number of active children of the parent of S in the original n-node tree.

We now have $\sum_{i:S_i \text{ is an "other" cluster}} |disc(S_i)| \leq k_o M + k_o + M \sum_{i:S_i \text{ is an "other" cluster}} a_{parent(S_i)} \leq k_o M + k_o + Mk$, since $\sum_v a_v \leq k$ and no node has two "other" children. Now $\sum_{\text{all } i} |disc(S_i)| \leq (k_s \cdot 1) + (k_t M) + (k_o M + k_o + Mk) \leq k + 2kM$. ■

Note. Via Lemma 4, we can guarantee that $M = 1$ and hence, by Lemma 5, that $D = 3K$.

Definition 10. *Let $W_0 = \lceil 10D \ln(\max\{K, 1/\varepsilon, 10\})/\varepsilon \rceil$.*

In the rest of this section we will prove the following theorem.

Theorem 2. *The tree Z produced by the algorithm is a k-node summary tree for T^w having (binary) entropy at least $OPT_k(T^w) - \varepsilon$, provided that W is chosen large enough that $W \geq D/K$ and that for $\eta = D/W$,*

$$\left(\frac{2}{\ln 2}\right) \eta \left(1 + \ln K - \ln \eta\right) \leq \varepsilon.$$

The least such W is at most W_0.

Let us first analyze the running time.

Theorem 3. *The running time of the algorithm is $O((K^3/\varepsilon) n \log(\max\{K, 1/\varepsilon\}))$.*

Proof. $\sum_{i=1}^{n} w'_i = W \leq W_0$. The running time of the exact algorithm is $O(K^2 nW)$ and W_0 is $O((K/\varepsilon) \log(\max\{K, 1/\varepsilon\}))$. ■

Definition 11. *For sequences $\langle p_1, p_2, ..., p_k\rangle$ and $\langle q_1, q_2, ..., q_k\rangle$ of the same length, k, of nonnegative reals summing to 1, let $H^e(p) = -\sum_{i=1}^{k} p_i \ln p_i$, where "$0 \ln 0$" is taken to be 0, and let $||p - q||_1 = \sum_{i=1}^{k} |p_i - q_i|$.*

To prove Theorem 2, we need a lemma, equation (55) in [Nau04], which is a quantitative version of the statement that almost identical probability distributions on $\{1,2,\ldots,k\}$ have almost identical entropy.

Lemma 6. *[Nau04, equation (55)] For $2 \le k \le K$, sequences $\langle p_1,p_2,\ldots,p_k\rangle$, $\langle q_1,q_2,\ldots,q_k\rangle$ of the same length of nonnegative reals summing to 1, and $\gamma \le k$ such that $||p-q||_1 \le \gamma$,*

$$|H^e(p)-H^e(q)| \le \gamma(1+\ln K-\ln\gamma).$$

We need a simple lemma whose proof uses Lemma 6. First we need a few definitions, which deal with two different k-node summary trees. (We will apply this lemma to the optimal k-node summary trees for $\langle w_1,w_2,\ldots,w_n\rangle$ and $\langle w'_1,w'_2,\ldots,w'_n\rangle$.) Definition 12 and Lemma 7 essentially state that for a fixed partition of $\{1,2,\ldots,n\}$ into k parts, the associated probability distributions defined by w and w' will have almost the same entropy, provided that $\eta = D/W$ is small. Quantitatively the lemma tells us how large W must be, in order to guarantee error less than ε.

Definition 12.

1. *Let $\langle S'_1,S'_2,\ldots,S'_k\rangle$ be a partition of $\{1,2,\ldots,n\}$ given by a k-node summary tree. Let $\hat{s}_j = \sum_{i\in S'_j} w_i$ and let $\hat{p}_j = \hat{s}_j/W$. Analogously, for the w''s, let $s'_j = \sum_{i\in S'_j} w'_i$ and $p'_j = s'_j/W$.*
2. *Let $\langle S_1,S_2,\ldots,S_k\rangle$ be a second partition of $\{1,2,\ldots,n\}$ given by a k-node summary tree. Just as above, let $s_j = \sum_{i\in S_j} w_i$ and let $p_j = s_j/W$, and analogously let $\bar{s}_j = \sum_{i\in S_j} w'_i$ and $\bar{p}_j = \bar{s}_j/W$.*
3. *Last, let $\Delta = \eta(1+\ln K-\ln\eta)$, where recall from Theorem 2 that $\eta = D/W$.*

Now we are ready for our lemma.

Lemma 7. *1.*

$$|H^e(\hat{p})-H^e(p')| \le \Delta \tag{3}$$

and

2.

$$|H^e(p)-H^e(\bar{p})| \le \Delta. \tag{4}$$

Proof. We have

$$||\hat{p}-p'||_1 = \sum_{j=1}^{k} |\hat{p}_j - p'_j||$$

$$= \frac{1}{W}\sum_{j=1}^{k}\left|\left(\sum_{i\in S'_j} w_i\right)-\left(\sum_{i\in S'_j} w'_i\right)\right|$$

$$= \frac{1}{W}|disc(\langle S'_1,S'_2,\ldots,S'_k\rangle)| \le \frac{D}{W} = \eta.$$

By Lemma 6,

$$|H^e(\hat{p})-H^e(p')| \le \eta\,(1+\ln K-\ln\eta) = \Delta.$$

Similarly,

$$||p-\bar{p}||_1 = \sum_{j=1}^{k} |p_j - \bar{p}_j||$$

$$= \frac{1}{W}\sum_{j=1}^{k}\left|\left(\sum_{i\in S_j} w_i\right)-\left(\sum_{i\in S_j} w'_i\right)\right|$$

$$\le \frac{1}{W}|disc(\langle S_1,S_2,\ldots,S_k\rangle)| \le \frac{D}{W} = \eta.$$

By Lemma 6,

$$|H^e(p)-H^e(\bar{p})| \le \eta\,(1+\ln K-\ln\eta) = \Delta. \blacksquare$$

Here is the proof of Theorem 2.

Proof. Let $\langle S'_1,S'_2,\ldots,S'_k\rangle$ be the partition of $\{1,2,\ldots,n\}$ defined by a k-node summary tree of maximum entropy for weights $\langle w'_1,w'_2,\ldots,w'_n\rangle$. Similarly, let $\langle S_1,S_2,\ldots,S_k\rangle$ be the partition of $\{1,2,\ldots,n\}$ defined by a k-node summary tree of maximum entropy for weights $\langle w_1,w_2,\ldots,w_n\rangle$. Equation (4) shows that there is a k-node summary tree for $T^{w'}$ (using $\langle S_1,S_2,\ldots,S_k\rangle$) of entropy at least $H^e(\bar{p}) \ge H^e(p)-\Delta = OPT^e_k(T^w)-\Delta$, where $OPT^e_k(T^w) = (\ln 2)OPT_k(T^w)$ is the optimal value of H^e over k-node summary trees of T^w. Therefore

$$OPT^e_k(T^{w'}) \ge H^e(\bar{p}) \ge OPT^e_k(T^w)-\Delta. \tag{5}$$

It follows that the entropy $H^e(T')$ of the output tree (which has weights derived from w, not w') satisfies $H^e(T) = H^e(\hat{p}) \ge H^e(p')-\Delta$ (by (3)), which equals $OPT^e_k(T^{w'})-\Delta \ge (OPT^e_k(T^w)-\Delta)-\Delta$ (by (5)), which equals $OPT^e_k(T^w)-2\Delta$. Converting now from natural to binary entropy, we have $H_w(T') \ge OPT_k(T^w)-\left(\frac{2}{\ln 2}\right)\Delta$. Now it is a simple matter to choose W to be the least positive integer at least D/k (so that $\eta = D/W \le k$) such that

$$\left(\frac{2}{\ln 2}\right)\frac{D}{W}\left(1+\ln K+\ln\frac{W}{D}\right) \le \varepsilon.$$

The reader can verify that the optimal W satisfies $W \le W_0$. $\blacksquare$

Since $g(x) = (D/x)(1+\ln K+\ln(x/D))$ is decreasing on $(D/K,\infty)$, one can use binary search on $[\lceil D/K\rceil, W_0]$ to find the smallest integer W in that interval with $g(W) \le \varepsilon$.

9.2. Proof of lemma 2

Proof. It is clear that for $l \ge 2$, $g_v(l,1,s_{v_1}+s_{v_2}+\cdots+s_{v_l}) = 0$ and $g_v(l,1,w) = -\infty$ for all other w, $-1 \le w \le W$.

Now suppose $l,k \ge 2$. For part 2., an optimal k-node summary tree for the union of the first l subtrees and having no "other" node must consist of an optimal summary tree for the first $l-1$ children, which has no "other" node, and having some number k_1 of nodes, together with an optimal $(k-k_1)$-node summary tree for the subtree rooted at the lth child. (The two summary trees must be optimal by equation (2), since otherwise the final tree would not be optimal.)

For part 3., consider an optimal summary tree for the union of the subtrees rooted at the first l children, having an "other" node of weight $w \geq 0$. Let Z be the set represented by the "other" node. By the definition of a summary tree, either $Z \cap V(T_{v_l}) = \emptyset$ (which is case (a)), or $Z = V(T_{v_l})$ (which is case (b)), or $Z \underset{\neq}{\supset} V(T_{v_l})$ (which is the most complicated case, (c)).

In case (a), we must have a summary tree for the first $l-1$ nodes having some number k_1, $1 \leq k_1 \leq k-1$, of nodes (all such possible values for k_1 being valid), together with a summary tree on $k-k_1$ nodes having no "other" node for T_{v_l}. (If $k-k_1 = 1$, the summary tree for T_{v_l} may or may not contain an "other" node; it doesn't matter.) Both summary trees must be of maximum entropy, as otherwise, by equation (2), the final tree would not have maximum entropy. That the formula given in part (a) is correct follows from the computation of the entropy of the resulting tree.

In case (b), we must have a summary tree for the first $l-1$ children, which has no "other" node, which is combined with a 1-node summary tree for T_{v_l} which has an "other" node of weight $w = s_{v_l}$. The formula is correct since it simply gives the entropy of the resulting summary tree.

Case (c) is tricky. We had a summary tree of the union of the subtrees rooted at the first $l-1$ children, one cluster of which was "other," plus one tree, all of whose nodes are together in one "other" cluster. We have to "merge" the "other" node of the summary tree for the first $l-1$ children with the set $V(T_{v_l})$, to get an enlarged "other" node (and a k-node summary tree). It follows that the summary tree of the first $l-1$ nodes must have had k nodes.

Computing the entropy of the new tree is not trivial. Specifically, let $M = s_{v_1} + \cdots + s_{v_{l-1}}$ be the sum of the weights of all nodes in the k-node summary tree, including the one cluster labeled "other." Let $w_L = w - s_{v_l} \geq 0$ be the sum of the weights of all nodes in the "other" cluster in the k-node summary tree. Let $w_R = s_{v_l}$. Let H be the entropy of that collection of k trees.

The surprising lemma that makes dynamic programming feasible is that the entropy of the collection of k sets in which the set $V(T_{v_l})$ is merged with the "other" cluster of the first $l-1$ children, to get an enlarged "other" cluster, is given by a function of H, M, w_L and w_R alone (and doesn't depend, for example, on the numbers of nodes in individual clusters, or their weights). However, we omit the detailed calculation justifying the value of H'. ■

9.3. A bad example for greedy

An interesting question is, how much smaller than the optimal entropy can the entropy obtained from the greedy heuristic be? Here we give an example for which the heuristic returns a 4-node summary tree of entropy only $2/3$ that of the optimal 4-node summary tree.

Let T be a tree on $\{1,2,...,7\}$, with node 1 as the root, having edges $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{2,5\}$, $\{3,6\}$, and $\{4,7\}$. Nodes 1, 2, 4, and 5 have weight 0. Nodes 3 and 6 have weight 1, and node 7 has weight 2. Sorting the children of node 1 into nondecreasing order by size gives $\langle 2,3,4 \rangle$. However, there is a 4-node summary tree of entropy 1.5 which has clusters $\{1\}$, $\{3\}$, $\{6\}$, $\{2,4,5,7\}$. The entropy associated with this tree is $H(1/4,1/4,2/4) = 2 \cdot (1/4)\lg 4 + (1/2)\lg 2 = 1.5$. The greedy algorithm produces the following vectors for $K = 4$:

1. $\text{elist}_2 = \langle 0,0,-\infty,-\infty \rangle$.
2. $\text{elist}_3 = \langle 0,1,-\infty,-\infty \rangle$.
3. $L_3 = \langle 0,0,1,1 \rangle$.
4. $\text{elist}_4 = \langle 0,0,-\infty,-\infty \rangle$.
5. $L_4 = \langle 0,1,1,1.5 \rangle$.
6. Final output entropy vector, after attaching the root: $\langle 0,0,1,1 \rangle$.

Hence, for $k = 4$, the optimal algorithm obtains 1.5 bits of entropy, as contrasted with the 1 bit obtained by the heuristic, thereby obtaining $2/3$ of the available entropy.

However, we have no example for which greedy obtains only $2/3$ of the optimal entropy, *when the optimal entropy goes to infinity*. Nor do we know if there is any fixed positive lower bound on the ratio between the entropy obtained by greedy and the optimal entropy, the so-called *performance ratio*.