

Visualizing the structure of large trees

Burcu Aydın*

*HP Laboratories
1501 Page Mill Rd MS 1040 Palo Alto CA 94304
e-mail: aydin@hp.com*

Gábor Pataki

*UNC at Chapel Hill
307 Hanes Hall CB 3260 Chapel Hill NC 27599
e-mail: pataki@email.unc.edu*

Haonan Wang[†]

*Colorado State University
216 Statistics Building Fort Collins CO 80523
e-mail: wanghn@stat.colostate.edu*

Alim Ladha and Elizabeth Bullitt[‡]

*UNC at Chapel Hill
CB 7062 Department of Neurosurgery Chapel Hill NC 27599
e-mail: ladha@email.unc.edu; bullitt@med.unc.edu*

and

J.S. Marron[§]

*UNC at Chapel Hill
352 Hanes Hall CB 3260 Chapel Hill NC 27599
e-mail: marron@email.unc.edu*

Abstract: This study introduces a new method of visualizing complex tree structured objects. The usefulness of this method is illustrated in the context of detecting unexpected features in a data set of very large trees. The major contribution is a novel two-dimensional graphical representation of each tree, with a covariate coded by color.

The motivating data set contains three dimensional representations of brain artery systems of 105 subjects. Due to inaccuracies inherent in the medical imaging techniques, issues with the reconstruction algorithms and inconsistencies introduced by manual adjustment, various discrepancies are present in the data. The proposed representation enables quick visual detection of the most common discrepancies. For our driving example, this tool led to the modification of 10% of the artery trees and deletion of 6.7%.

*Partially supported by NSF grants DMS-0606577 and DMS-0854908, and NIH Grant RFA-ES-04-008.

[†]Partially supported by NSF grants DMS-0706761 and DMS-0854903.

[‡]Partially supported by NIH grants R01EB000219-NIH-NIBIB and R01 CA124608-NIH-NCI.

[§]Partially supported by NSF grants DMS-0606577 and DMS-0854908, and NIH Grant RFA-ES-04-008.

The benefits of our cleaning method are demonstrated through a statistical hypothesis test on the effects of aging on vessel structure. The data cleaning resulted in improved significance levels. Our second example analyses brain artery images of healthy patients and patients with brain tumor. Our visualization can identify tumor patients.

AMS 2000 subject classifications: Primary 62-09; secondary 62-07.

Keywords and phrases: Visualization, data cleaning, tree structure, descendant-level view, brain arteries.

Received June 2010.

1. Introduction

Most real life data sets contain a variety of challenging features, which can be noise artifacts or other kinds of discrepancies. Elimination of these artifacts may result in sharper statistical results. Relationships that were previously obscured may become more clear.

Our motivation comes from a data set of brain artery systems of 105 subjects collected by the CASILAB (casilab.med.unc.edu). The extraction of this data set from raw Magnetic Resonance Angiography (MRA) images are summarized in Section 2, and further details can be found in Aylward and Bullitt [2].

An earlier version of this data set consisting of 73 data points (**Data Set 1**) was used to statistically analyze the effect of aging in brain vessel structure in Wang and Marron [10] and Aydın et al. [1]. In the latter paper, the rich three dimensional structure of the vessel systems are summarized by binary trees which only keep connectivity information. The aim was to strip the features other than branching from the data and obtain a simplified representation to study the effect of aging on the branching structure of the vessels. Figure 1 shows the 3 – D image of a brain vessel system and one of the binary trees extracted from it.

Since the original analysis, 34 more subjects were added to the study and two low quality cases were deleted. Careful anatomical examination of each data tree revealed some errors in the flow direction of vessels and revealed that starting points were arbitrary depending on the head position in the scanner. The 3 dimensional trees (as in the left panel of Figure 1) in **Data Set 1** were constructed by an automatic vessel connection algorithm (see Aylward and Bullitt [2] for details) which occasionally resulted in anatomically incorrect connections. These issues were addressed, in a painstaking case by case fashion, by manual reconnection of tree components. The head position problem was addressed by starting each tree at the *circle of Willis* (well known, common component of human brains). When there were more than four trees, trees were combined to result in the best approximation of front, left, right, back flow systems. The resulting data set is called **Data Set 2**. As shown in Aydın et al. [1], the same statistical methods applied to this new data set reveal a remarkable improvement in significance levels. This demonstrates the value of improved data quality.

However, even after this cleaning process, many discrepancies are known to remain in the data set. The details of 3 separate problems that we found in the

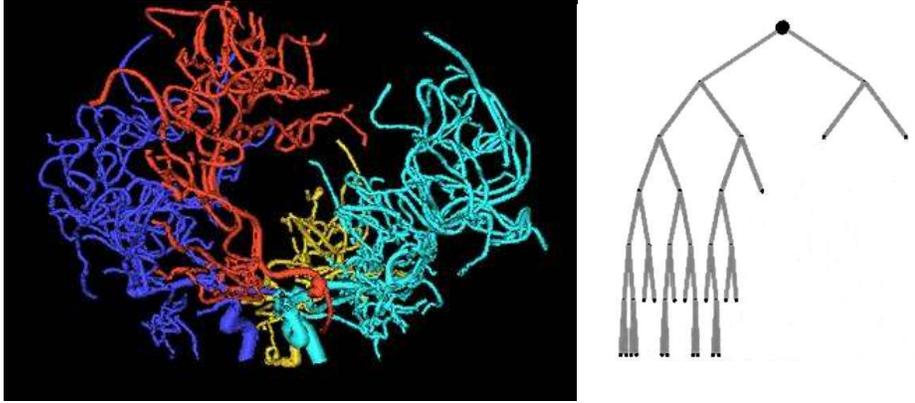


FIG 1. On the left: Reconstructed set of trees of brain arteries. The colors indicate regions of the brain: Back (gold), Right (blue), Front (red), Left (cyan). On the right: Binary tree obtained from the back (gold tree) of the same subject. Only branching information is retained.

data are explained in Section 4. Some of these discrepancies are fundamental in a way that they may change the observed structure of the vessel system. Elimination of as many of these problems as possible will improve the quality of a statistical analysis.

Another point of interest for this data set is the effect of various diseases on the vessels. Bullitt et al. [3] studies these effects in detail. They identify three patterns of anomalies in brain vessels and introduce metrics that can be used to identify affected vessels. As a second example we focus here on *Type III* pattern. Type III abnormalities are apparent in malignant brain tumors. These vessels can be any length, and can be straight or curved. The underlying pattern is that, near tumors, they exhibit high frequency coils.

The Type III anomalies associated with a tumor may be visible before the tumor is large enough to be noticed in screenings. Therefore the ability to recognize these is a valuable input for a clinician. Bullitt et al. [3] uses the term *tortuosity* to describe the coiling, twisting pattern of Type III. They introduce the *Sum of Angles Metric (SOAM)* to measure tortuosity of vessels.

The full 3 dimensional tree structures, e.g. as shown in the left panel of Figure 1, contain a large amount of visual information. This makes it hard to see and understand the relevant problems with the data. The right panel of Figure 1 shows a simplified, topology only structure, which enables focusing on purely topological aspects. However, this representation only works to node level 10 to 11 where there is not space to display more nodes. A major contribution of this paper is, due to the graphical representation that enables focusing on the important aspects of each data tree, we are able to obtain a simplified view that can be quickly inspected. This view carries the important information necessary to spot problems and eliminates others.

One problem with the 3 – *D* representation is that, the amount of details present in the set is very high. Examining this set thoroughly requires checking

the linkage and position of each branch in each subject separately, an extremely tedious job to process manually. It is important to reduce the level of complication without losing the aspects necessary to track down discrepancies.

Graphically representing each data tree can be a powerful method to both understand the data and discover any problems. Carefully designing a visual representation of the data provides efficient visual inspection of each of these instances and may enable these discoveries without any further diagnostics. The challenge is to clearly display the important aspects of the data and eliminate the details that will not help with the diagnostics.

In this paper we propose a visualization method that enables the discovery of discrepancies with ease through a concise summary which eliminates irrelevant details. The definition of *concise summary* is agreed upon after carefully analyzing the structure of the problem, and the nature of common problems in it. For our first example where we aim to flag structural discrepancies, we recognized that the key aspects are:

- Number of descendants of a node in the tree,
- Level of a node,
- Vessel thickness.

For example, misconnections are easily seen as thick vessels appearing at a low level or as a low level node with many descendants. For our second example, along with structural properties of each node (number of descendants and level), the tortuosity measure of each node is the key to discover possible tumors.

Our visualization could easily be modified to display additional properties using node size, node shape, edge color and edge thickness, depending on the nature of the data set to be displayed. We expect our method to be useful for displaying any tree set where branching structure is of interest. Example areas include social networks, evolutionary biology, and medical imaging. For example, in the latter area, there are many branching structures in the human body including vasculature, airways, lymph systems, etc.

Effective ways of visualizing trees have been previously proposed in the literature. A widely studied very special case is phylogenetic trees. Letunic and Bork [5] and Huson et al. [4] are two recent studies, which include good literature surveys on the subject for the further interested reader. The main focus of phylogenetic tree visualization studies is to clearly represent the relationships between a given set of species, where each species is represented as a leaf on the tree. The aim is to show the taxonomical hierarchies between species which is revealed through the structure of the parent nodes.

These trees start out from an identified set of leaf nodes (species) and aim to aid researchers in studying the possible evolutionary relationships through the constructed phylogenetic tree's structure. A key aspect for phylogenetic trees is that, the leaf set for all the trees in one population is fixed (these are current species). What changes from one tree to another is the structure above the leaves. This structure represents the predicted ancestry relationship between existing species.

Our context is much more general than phylogenetic tree because the leaves do not correspond to each other across the members of a population. In particular, the generating mechanisms of the structure lead to differing leaf sets. Therefore the visualization methods developed to construct phylogenetic trees are inadequate for our more general set of data objects.

In the context of very large trees, Nguyen and Huang [6] and Nguyen and Huang [7] studied methods to visualize these in 2 dimensional space. They use a concept called *enclosure* to partition the entire display space into a collection of local regions that are assigned to each node of a tree. Their approach is different from ours because they are dealing with very large trees, up to 50000 nodes. Because our trees are comparatively moderately sized (100 – 200 nodes), our approach enables us to focus more on tree structure rather than on space optimization.

Search trees generated by finite domain constraint programs have also been investigated in the literature. Simonis and Aggoun [8] did a visualization study of such trees, using a software tool they developed to debug algorithms through tree visualizations. Though interesting, the nature of these trees are very different from our data set, and methods to analyze them are not applicable to our situation.

Finally, a recently popular approach is *tree-maps*. Tree-maps transform the traditional branched tree view into a rectangle divided into sub-rectangles, arranged according to some properties of interest of the nodes. Shneiderman [9] proposed the tree-map idea, and numerous variations have been studied in the literature. This approach is useful with categorical data where a parent node represents an attribute common to all its children. This approach again is not suitable for our data set.

The organization of this paper is as follows: Section 2 gives a description of how the binary trees used for statistical analysis are produced from the $3 - D$ representations as shown in Figure 1. Section 3 develops the details of the proposed visualization method. Section 4 indicates some problems that commonly arise within the data set. Section 5 explains how the visualization method is used to identify these problems. The results of the cleaning process together with a comparative statistical analysis is also given in this section.

2. Extraction of binary trees from the raw data

Following Aydın et al. [1], the way the binary trees are extracted is as follows: For each of the instances (brain scans), the back, left, right and front regions are handled separately. Each of these subsystems usually consist of one main (root) vessel entering the brain from below, and splitting into smaller branches to feed that region of the brain. The portion of the root vessel until a branch splits off is taken as the root node, and the two vessels that take place after the split are the left and right children nodes of the root. The same procedure is applied at each juncture point. In the end, a binary tree is obtained, where each vessel trunk between two split points in the original structure corresponds to a node.

An issue is, whether a vessel splits into three or more branches at a single point. In this data set splitting into more than two branches is very rare, and these occurrences do not carry any important implications on structure. Therefore keeping the simple binary structure seems more important than capturing these rare occurrences. In the cases that this happens, one of the child vessels is arbitrarily selected as the first one to split off, and the binary tree is created accordingly.

In some of the instances, there can be two root vessels feeding one region. In these cases, for simplicity, one phantom root node is added and the roots of these two trees are connected to the phantom root node as children, so that a single binary tree is obtained for each instance in the sub-populations. These binary trees are called *component trees*.

3. Visualization

An example of our visualization can be seen in Figure 2. Each node is represented with a colored dot and each parent-child couple is connected with a line segment, revealing the connectivity. An important indicator of a node's significance within the structure is the number of descendants that it has. The nodes with the larger number of descendant nodes are more central, and in our case, more crucial to the blood flow, compared to the ones with a smaller number of descendants. For

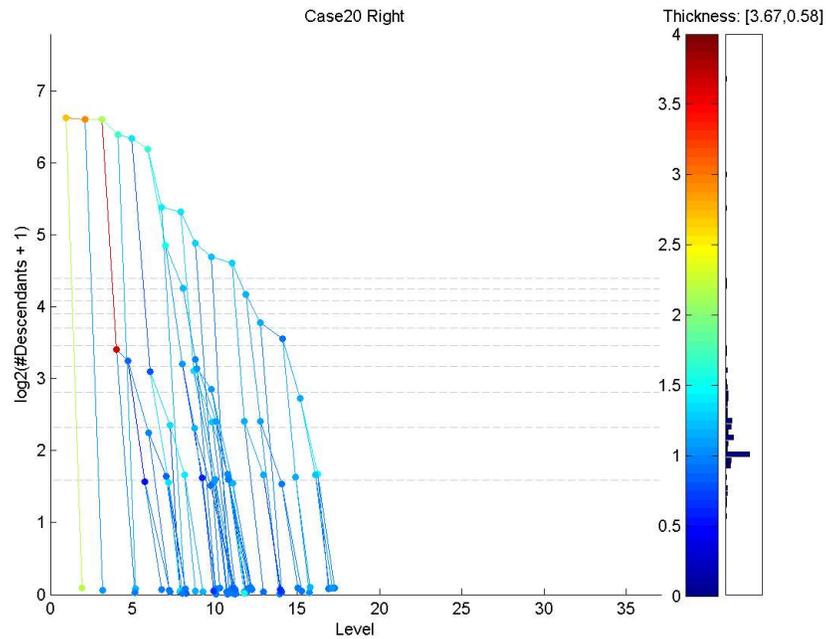


FIG 2. A visual display of the right component tree of subject 60, Data Set 2.

this reason we choose to use the y-coordinate to show the number of descendants of each node. For compact summarization the logarithm (base 2) is used on this axis. The x-axis shows the level of the node in the binary tree. The level signifies how close each node is to the root node. Because of this axis selection, the visualization is called the *D-L view* (Descendant-Level view). This view can essentially be seen as a scatter plot of nodes connected to each other with respect to their parent-child relations.

In the lower levels, many nodes have the same coordinates (number of descendants and level). To avoid the over-plotting of these nodes, some jittering (adding small amounts of random displacement for improved visual separation) is applied at the lower levels.

In Figure 2, color is used to code thickness. Because the thickness of the vessel segments is important to understand tree structures and potential errors, the median thickness is coded by color when the aim is to identify structural problems. This choice was made because the thickness measure of vessels in this data set is a major indicator of data problems: nodes thicker than the normal range or their parents usually point to a tracking or recording problem. In addition, color coding enables quick visual scans and brings out the problematic nodes even in crowded areas. For other kinds of data sets it is possible to reserve color for another important variable, as in our brain tumor example.

In the brain vessel data set, thickness rarely exceeds 4 millimeters, so the range of thickness used is $[0, 4]$. Thicker nodes are put in the top 4mm bin. This range is linearly projected onto a color map which consists of 100 shades, or bins, shown as the color bar on the right. The thicknesses close to 4mm correspond to dark red shades. As vessels thin, the color follows a range through yellow, green, to a dark shade of blue. The distribution of thicknesses is summarized in a bar chart, indicating the counts of the nodes that fall into the range of each bin, displayed on the right hand side of the figure. The numbers displayed on the right top corner show the thickness range of the nodes for that particular data tree, i.e. in Figure 2 the thinnest node has a thickness of 0.58mm.

4. Identification of common structural problems

Based on anatomical knowledge and experience, several types of discrepancies that may exist in the data have been identified. Note that some natural noise exists in this data along with tracking and labeling errors listed below. An important aim of the D-L view was to design a clear representation of the data such that these errors can be told apart from the natural noise. In Figure 2, on level 4, a vessel segment thicker than its parent exists, which is not a regular occurrence. However, the thickness difference of that segment with its parent is within measurement error range, so this instance is not flagged for checking.

The **Data Set 2** is a result of a different, more anatomically based, case by case clean up process and the addition of 34 more cases. That clean-up process aimed to correct the cut off point problems of the vessel systems, by manually going through each instance and examining them. This study is meant to develop

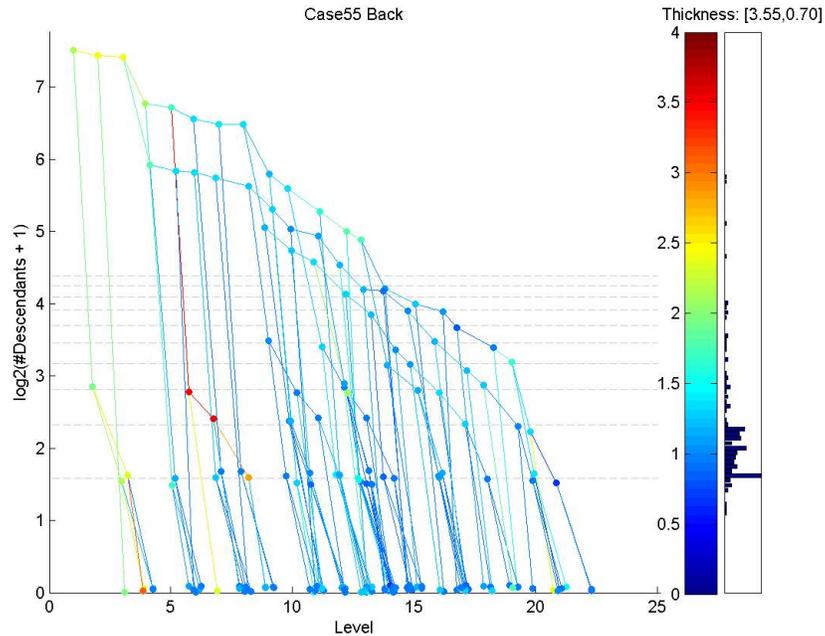


FIG 3. The D-L view of the back component tree of subject 55. Red/yellow subtree starting from level 6 is a possible misconnection.

a visual method to diagnose a wider range of problems without scrutinizing the raw data form as shown in Figure 1, which is loaded with information and thus is hard to examine. The methods explained here are applied to Data Set 2, as presented in Section 5, to obtain Data Set 3.

We have identified the following major kinds of discrepancies that can be corrected through inspection:

- **Misconnections:** Some vessels that are not anatomically connected appear connected in the representation seen in left panel of Figure 1, due to being close to each other relative to the accuracy of the MRA image slices. This error may result in misinterpretation of the blood flow of direction, and one of these vessels is seen as an extension of the other in the model. Normally, the vessel trunks that are connected to each other are expected to have similar thicknesses, and this thickness should generally decrease as one goes from root nodes to the leaves. A sudden jump in thickness is an indication of misconnection, i.e. when a thinner parent node has a much thicker sub-tree descending from it. Figure 3 shows an example of this situation. Notice that the red/yellow subtree starting from level 6 is much thicker than its parent vessel.
- **Starting Point Problems:** Determining the point where a vessel enters the brain highly depends on the cutoff point of the MRA images (depend-

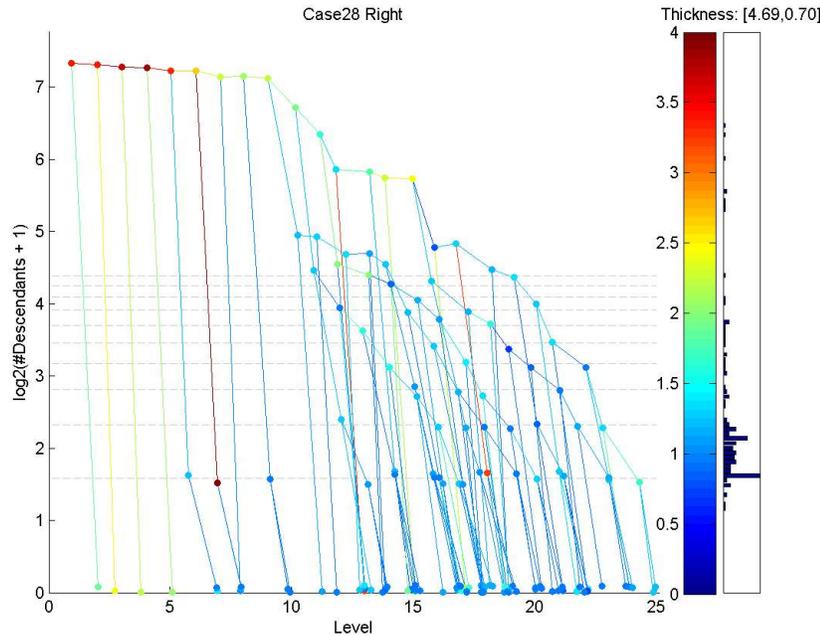


FIG 4. The D-L view of the right component tree of subject 28. The irregularities close to the root node indicate a possible starting point problem.

ing on head position in the scanner). In some cases, this cutoff point is mistakenly taken at a too low or too high level. It is also possible to mark a child of a root as the root node. Figure 4 shows a tree with a possible starting point problem. The initial series of thick red nodes suggest that the MRA starting point was taken too low.

- **Veins:** The tracking system used to obtain the data trees is intended to record only arteries in the brain. However, in some cases, veins that run very close to an artery are mistakenly identified as an extension of that artery. These veins are usually thicker than the parent artery, and show up as red leaf nodes on the visualization. An example of this kind of problem is displayed in Figure 5.

5. Solutions and results

5.1. Discovering structural discrepancies

Through a careful inspection of the visual displays of all points in the data set, problematic instances are marked to be reviewed again. The numbers of instances marked for review for each sub-population are given in the top row of Table 1.

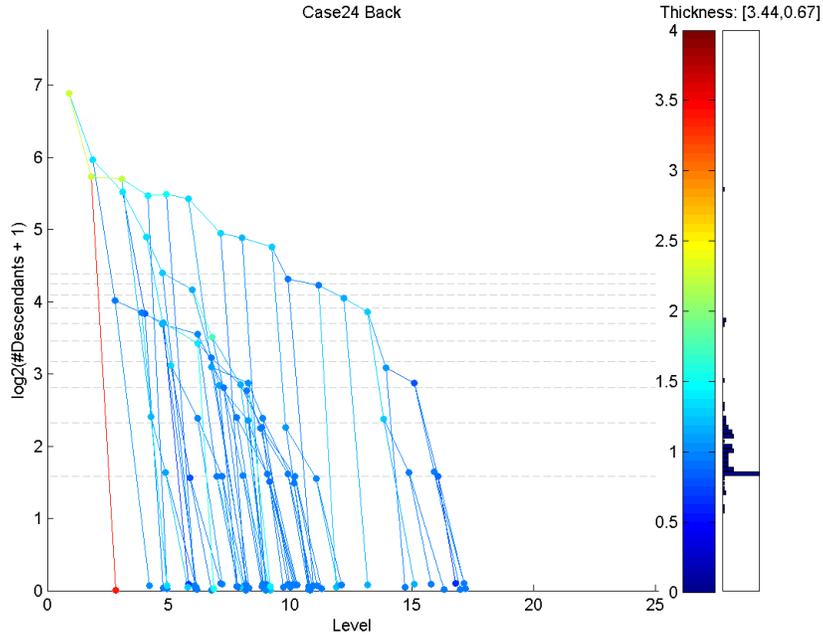


FIG 5. The D-L view of the back component tree of subject 24. The red leaf on the third level is identified as a possible vein.

TABLE 1
 Numbers of instances marked for reviewing (top row) and numbers of instances manually modified after raw data inspection (bottom row) for each kind of problem and for each tree location

Misconnection				Starting Point				Vein			
B	L	R	F	B	L	R	F	B	L	R	F
7	3	4	2	3	6	4	3	18	16	23	9
4	2	3	0	2	6	4	3	4	1	5	1

Note that the 4 sub-populations contain a total of 420 data trees. The process resulted in spotting 98 of them (24%) as potentially problematic instances that may significantly benefit from manual review of the 3 – D raw data.

As a result of a careful study of the instances flagged, 7 cases were identified as severely problematic (consistent with major errors in the scanning process), and were excluded from the data set. For the remaining cases, the bottom row of Table 1 shows the number of instances that were manually modified out of each problem group.

Of the remaining marked component trees, 35 were selected for manual modification based on inspection of the raw data. The artifacts in the remaining 56 instances were manually determined to be consistent with the natural noise.

While our visual diagnostic found many errors among the tree data, it is not perfect. In particular, one of the marked trees revealed an error different than flagged by the diagnostic.

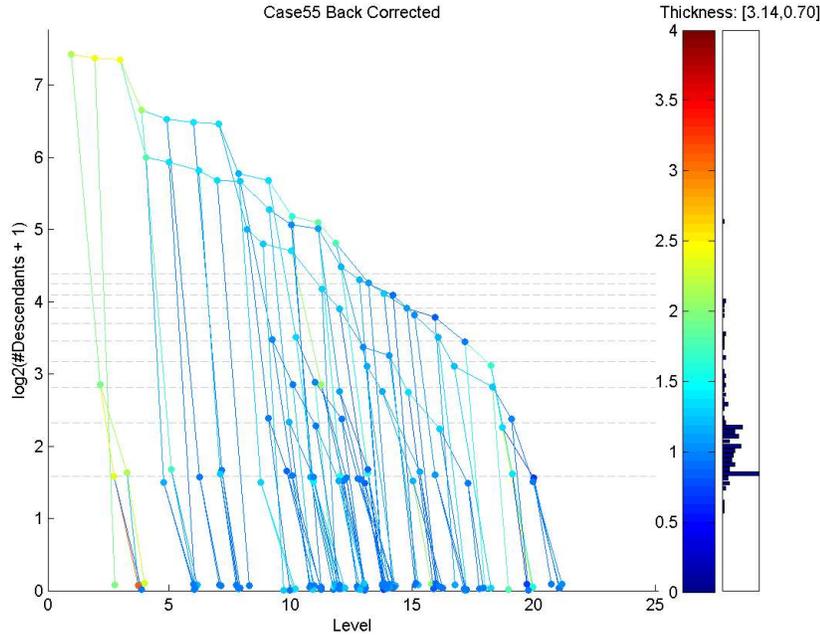


FIG 6. The D-L view of the back component tree of subject 55 (Figure 3) after correction. The misconnected subtree is deleted.

Next, our visualization of the cleaned version of each of the above three example cases (with discrepancies) will be studied.

Figure 6 shows the corrected version of Case 55, back component tree (from Figure 3). Notice that the red/yellow subtree that was incorrectly attached to this tree has been removed.

Figure 7 shows the result of fixing the starting point problem with the Case 28 right component tree. Notice that the irregularity close to the root node (a series of dark red nodes) has been corrected by changing the cut-off level to a higher point.

Figure 8 shows that, after revision of subject 24's back component tree, the red leaf node seen in Figure 5 was indeed identified as a vein and was removed from the data tree.

All of these modifications in the data resulted in **Data Set 3**. In particular, 7 high-problem cases have been removed from **Data Set 2** and the 35 component trees have been corrected.

An important question regarding the clean up study is that: does the elimination of several problems improve the data set in terms of age effect? In other words, is the age effect more pronounced now or not? Wang and Marron [10] and Aydin et al. [1] have previously used a statistical analysis tool called *tree-line analysis* to measure the effect of aging on brain vessel structure in a smaller

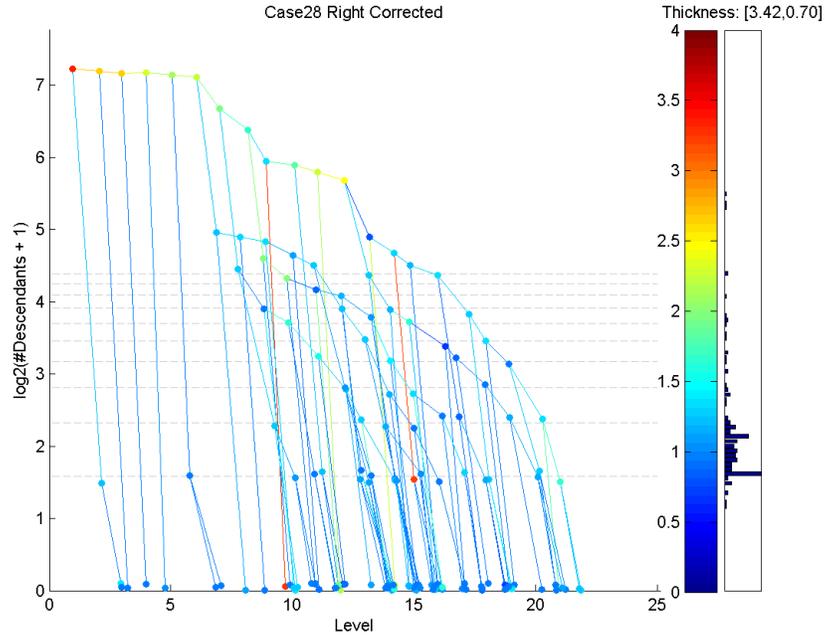


FIG 7. The D-L view of the right component tree of subject 28 (Figure 4), corrected by changing the cut-off point of the root vessel.

TABLE 2

The slope p -values of first principal components obtained using the tree-line tool. Columns compare the Data Sets 2 and 3. Smaller values point to greater significance level

	Set 3	Set 2
Back	0.0318	0.0715
Front	0.0296	0.0436
Right	0.0685	0.0743
Left	0.0916	0.0493

data set (Data Set 1). The same tool is used here to compare Data Set 2 and 3 in terms of age effect.

Table 2 summarizes the slope p -value comparisons of two data sets (2&3) using the tree-line tool. It displays the the slope p -values obtained from first principal components against the age variable. Smaller p -values point to a greater statistical significance level, that is, a tighter relation between age and projected tree sizes. According to the findings of this analysis, the significance level increases (p -value decreases) in 3 out of 4 sub-populations with the cleaning. This means the age effect on brain vessel structure gets more pronounced after the cleaning process, in accordance with the expectations.

Although the cleaning process generated an improvement on age-projection size relation, the change is not dramatic. It is clear that the visualization process

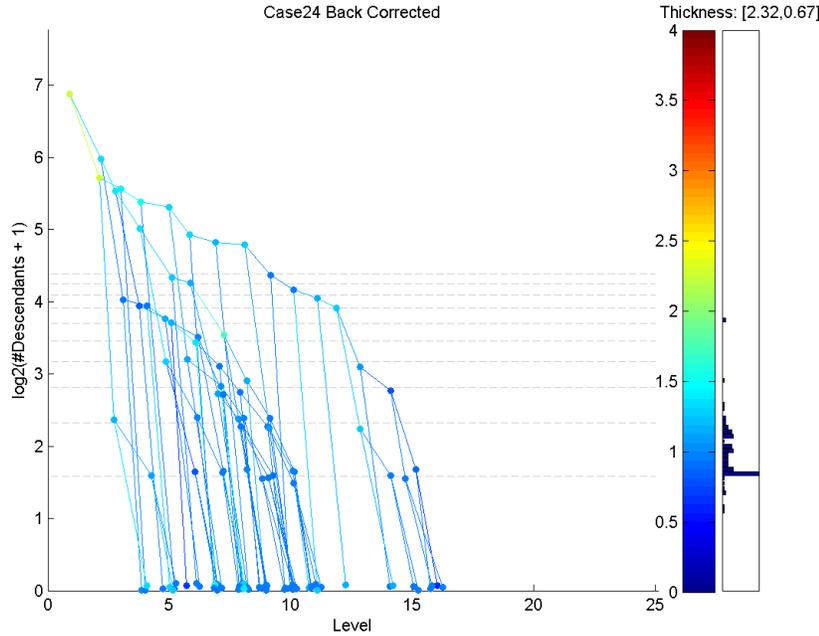


FIG 8. The D-L view of the back component tree of subject 24 (Figure 5), corrected by the deletion of a vein.

resulted in the correction of many instances and thus provided a more reliable data set. However, most of these corrections are local compared to the sizes of whole trees within the sub-populations. The effect of aging is related to the general trend of *branchiness* within those trees, and this trend did not dramatically change with our cleaning: data trees with many branches still have a lot of branches and vice versa.

The visualization method is not claimed to detect all possible problems that exist in the data sets. It relies on the irregular thickness differences of seemingly adjacent nodes. It is possible that there are still instances in the data set that contain for example misconnected vessels, but if those vessels wrongly connected have similar thicknesses, they will show up as normal branches on the visualization. However, there is no way of spotting those problems without using anatomical knowledge and the raw data, which would be an enormously time consuming task. The aim here is to provide a simple and quick new tool for finding problem instances.

5.2. Identifying brain tumor patients

In this section we will illustrate the use of our method to visually inspect the brain vessel system for a tumor marker (Type III tortuosity). In this example, we use color to code the SOAM measure for each node rather than thickness.

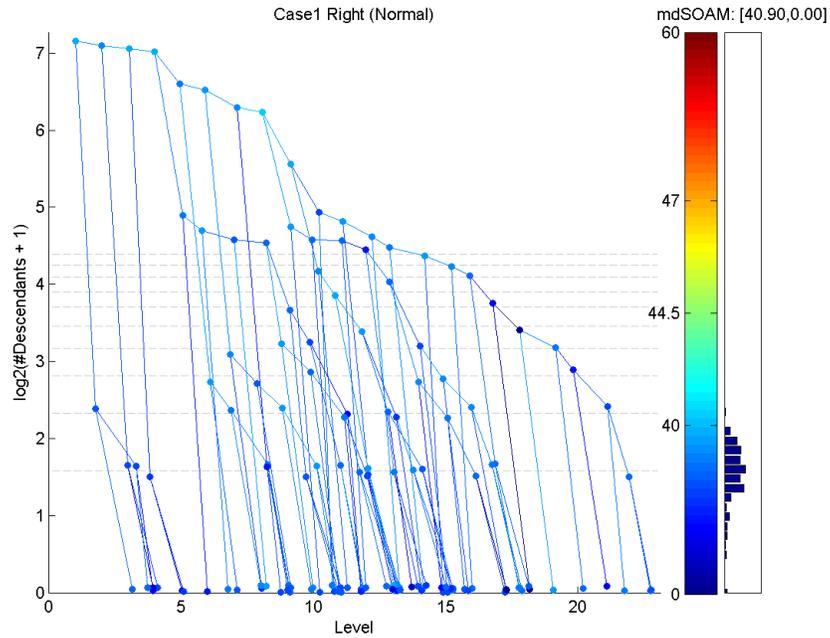


FIG 9. The D-L view of a healthy subject where color expresses the SOAM measure for each node.

Unlike the previous example, the coloring scheme is not linear. We have used a spline function to project the SOAM measure of each node onto the color range. Specifically, nodes with SOAM measures between 0 and 40 are linearly projected onto the blue range, between 41 – 44 are cyan, 45 – 47 are yellow and finally any measure at or above 48 are projected to the red region. This choice comes from the nature of this measure. The normal range seems to be between 0 and 40, and possibly up to 44 depending on the location. Higher measures, corresponding to yellow and red, should be watched out for.

Vessels around a tumor may not be adjacent in the D-L view, because physical proximity is not represented. Only physically attached vessels will appear as connected. Therefore the seemingly separate high tortuosity nodes seen in Figure 10 may be physically close, feeding a single tumor.

More example views of healthy subjects and tumor patients can be found in supplementary materials in the journal's web site. These shows similar lessons as represented in Figures 9 and 10.

5.3. Conclusions

In conclusion, the contribution of the visualization method is clear. In our structural diagnostics example, 43% of the instances that are marked for inspection revealed mistakes that it was appropriate to correct. As a result, the statistical

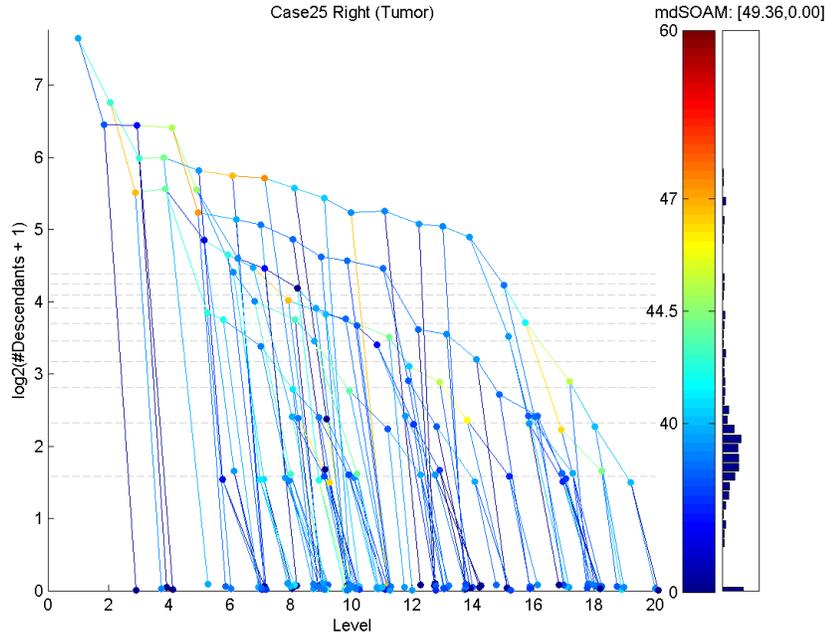


FIG 10. The D-L view of a subject with brain tumor where color codes the SOAM measure. The blue and cyan nodes have normal tortuosity levels. The high tortuosity nodes, indicated as red and yellow, imply these may be vessels feeding a tumor.

significance of age effect on vessel structure improved, pointing to a less noisy data. The method provides an effective diagnostic for data clean up in similar populations of binary trees. Our second example of comparison of healthy and tumorous brain views illustrates how the D-L view can be used to inspect vessel images for tumor markers.

References

- [1] AYDIN, B., PATAKI, G., WANG, H., BULLITT, E., MARRON, J.S. (2009) A Principal Component Analysis For Trees, *Annals of Applied Statistics*, 3, 1597–1615.
- [2] AYLWARD, S. AND BULLITT, E. (2002) Initialization, noise, singularities and scale in height ridge traversal for tubular object centerline extraction, *IEEE Transactions on Medical Imaging*, 21, 61–75.
- [3] BULLITT E, GERIG G, PIZER SM, AYLWARD S. (2003) Measuring tortuosity of the intracerebral vasculature from MRA images. *IEEE-TMI* 22:1163-1171.

- [4] HUSON, D., RICHTER, D., RAUSCH, C., DEZULIAN, T., FRANZ, M., RUPP, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees, *BMC Bioinformatics*, 8, 460.
- [5] LETUNIC, I., BORK, P. (2006) Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation, *Bioinformatics Applications Note*, 23-1, 127–128.
- [6] NGUYEN, Q., HUANG, M. (2002) A Space-Optimized Tree Visualization, *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, 85.
- [7] NGUYEN, Q., HUANG, M. (2007) Improvements of Space-Optimized Tree for Visualizing and Manipulating Very Large Hierarchies, *Pan-Sydney Area Workshop on Visual Information Processing (VIP2002), Sydney, Australia, Conferences in Research and Practice in Information Technology*, Vol. 22.
- [8] SIMONIS, H., AGGOUN, A. (2000) Search-Tree Visualization, *Lecture Notes In Computer Science*, Vol. 1870, 191–208.
- [9] SHNEIDERMAN, B. (1998) Tree Visualization With Tree-Maps: 2-D Space Filling Approach, *ACM Trans. Graph*, 11-1, 92–99.
- [10] WANG, H. AND MARRON, J. S. (2007) Object oriented data analysis: Sets of trees, *The Annals of Statistics*, 35, 1849–1873. [MR2363955](#)