# Lossy compression of unordered rooted trees

Romain Azaïs*, Jean-Baptiste Durand†, and Christophe Godin‡

*Inria team BIGS
Université de Lorraine
54 506 Vandœuvre-lès-Nancy, France
romain.azais@inria.fr

†Université Grenoble Alpes
Inria team MISTIS
38 041 Grenoble, France
jean-baptiste.durand@imag.fr

‡Inria team Virtual Plants
Université Montpellier 2
34 095 Montpellier, France
christophe.godin@inria.fr

**Abstract**

A classical compression method for trees is to exploit subtree repeats in the structure by representing them by directed acyclic graphs. We propose a lossy compression method that consists in computing a structure with high redundancy that approximates the initial data.

Trees are commonly used to represent hierarchical data appearing in computer science or in biology. Compression methods often take advantage of repeated substructures appearing in the tree (see the survey [1]). Directed Acyclic Graph (DAG) compression is a classical approach that exploits subtree repeats in the structure. However, it should be noted that trees without a high level of redundancy are often insufficiently compressed by this procedure. Self-nested trees are such that all their complete subtrees of a given height are isomorphic. The systematic repetition of subtrees gives them remarkable compression properties by this approach.

We address lossy compression for unordered trees. Loss can be acceptable for visual representation of scenes composed of plants, for example. Our method consists in computing the DAG version of a self-nested tree that closely approximates the tree to compress. A first approximation has been proposed in [2] in which the authors compute in polynomial time the Nearest Embedding Self-nested Tree (NEST) of the initial structure, namely the self-nested tree that minimizes the edit distance to the initial tree and that embeds it. We focus on the presentation of two new algorithms to find a self-nested structure that approximates the initial tree better than the NEST. These solutions rely on a technique to find the centroid of a forest of small height and may be computed in polynomial time for trees with bounded degree. We prove on a simulated dataset that the error rates of these lossy compression methods are always better than the loss involved in the previous algorithm (on average, we observe a substantial gain of around 20%), while the compression rates are equivalent.

## References

[1] S. Sakr, "XML compression techniques: A survey and comparison," *Journal of Computer and System Sciences*, vol. 75, no. 5, pp. 303 – 322, 2009.

[2] C. Godin and P. Ferraro, "Quantifying the degree of self-nestedness of trees. Application to the structural analysis of plants," *IEEE TCBB*, vol. 7, no. 4, pp. 688–703, Oct. 2010.