Some notes on the computer program used to compute pairwise k-RF scores of multiset-labeled trees

The computer program used to compute pairwise k-RF scores of all multiset-labeled (rooted) trees in an input text file, can be found in "k-RFmeasures.py". The code utilizes some functions defined in "functions.py" and has two propositional arguments "inputfile" and "k". By default, the program considers trees as unrooted; however, adding the optional argument "–r" or "––rooted" in the command line, changes the default. Therefore, the command lines "python3 k-RFmeasures.py inputfile k" and "python3 k-RFmeasures.py –r inputfile k" are used to compute pairwise k-RF measures of all multiset-labeled trees and all multiset-labeled rooted trees, respectively.

As mentioned above, an input file is one of the propositional arguments of the program. To prepare the file, one needs to represent each multiset-labeled (rooted) tree by its (directed) edges in separate lines and begin the representation by the phrase "tree (tree name)" (note that "tree name" can be omitted). The edges need to be written in a special format as shown for the rooted tree T in Figure 1 (the format for unrooted trees is similar to rooted trees). Some input file examples are "rooted.txt" and "unrooted.txt", which can be found in the GitHub link.

"5000trees.txt", "1000trees.txt", and "250trees.txt" are the other sample input files in the GitHub link, which consist of the data sets used for three experiments explained in the paper. The first two were used for the correlation analyses of k-RF measures with CASet$\cap$, DISC$\cap$, and GRF. Note that trees in "5000trees.txt" have the same label set; additionally, the first, second, third, fourth, and fifth 200 trees in "1000trees.txt" form 5 families of set-labeled trees such that trees inside a family have the same label set while trees across the families have different but overlapping label sets. In addition, "250trees.txt" was used for the clustering application of the k-RF measures, where the first, second, third, fourth, and fifth 50 trees form 5 families of set-labeled trees such that trees inside a family have the same label set while trees across the families have different but overlapping label sets.

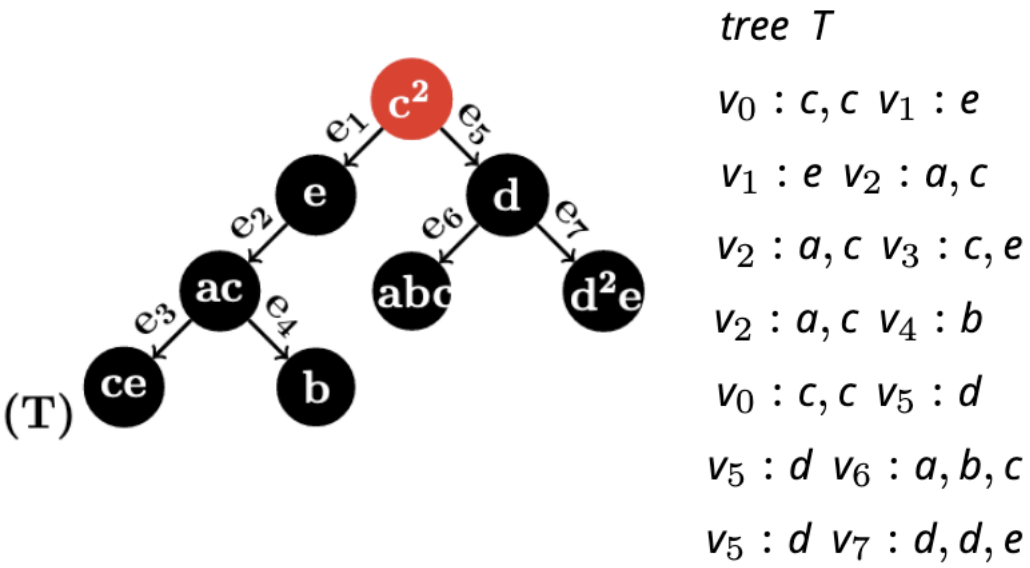

Figure 1: Representation of a multiset-labeled rooted tree

As illustrated in Figure 1, each edge $e = (u, v)$ is represented by the phrase "u:list of u's labels v:list of v's labels" in one row. Note that the list of u's labels is followed by an space.

In addition, for each multiset-labeled tree, labels of each node need to be listed in a fixed order throughout the representation of all edges. More precisely, if the node v is labeled by ab^2c, we first fix an order on $\{a, b, c\}$, such as $a < b < c$ and then represent the node's labels with the order as a, b, b, c in all edges with the node.