Detect Duplicate Nodes

Description

'Detect Duplicate Nodes' helps to find duplicate nodes in a network, by checking to see if any nodes in the network have labels that are very similar to each other.

The algorithm produces three results:

- A 'Merge Table' that can be used in conjunction with the Update Network by Merging Nodes algorithm to combine each set of duplicate nodes in a network into a single node.
- A human-readable report describing which nodes will be merged.
- A human-readable report describing which nodes will NOT be merged.

'Detect Duplicate Nodes' compares the label of every node with the label of every other node (but see the exception below), and uses the Jaro-Winkler text similarity to find the % similarity of the two labels. If the similarity is above the user-defined threshold (.95 by default), they are set to be merged in the Merge Table. If the similarity is above a second user-defined threshold (.85 by default), those two nodes will be recorded in the third output report which specified nodes that were close to being merged, but didn't quite make it. The purpose of this report is to let the user see if certain merging pairs should in fact be merged, so they can either adjust their merging threshold, or manually edit their Merge Table to include or disinclude nodes from the merging process.

Because there are roughly *n*n/2* comparisons between labels in a network of n nodes, the computational complexity of this algorithm can grow steeply as *n* increases. To reduce the runtime for large datasets, we have included an option to limit comparisons between nodes whose labels only share their first X number of letters (2 by default). You can increase this number to improve the speed of the algorithm, or decrease to improve the chances of catching duplicates where there is variation in the first X letters of two labels.

Pros & Cons

This algorithm will not detect 100% of the duplicates in a dataset in most cases, but can be a good first step to create a Merge Table which can then be checked by hand for 100% accuracy. It can also produce acceptable results on its own if 100% accuracy is not mandatory.

See Also

Usage Hints

Try running this algorithm once, then inspecting the "Nodes that will be merged" and "Noteworthy nodes that will NOT be merged" reports. If you agree with the recommendations in those reports, follow through by running Update Network by Merging Nodes. If too many nodes are being merged, or not enough, run this algorithm again with adjusted thresholds. Continue adjusting until you see the results you want. You may need to edit the Merge Table by hand for 100% accurate merging.