

Network Workbench

A Workbench for Network Scientists

Tutorial for Sunbelt 2008

Russell Duhon

Computer Science

Indiana University - Bloomington

rduhon@indiana.edu

Learn more about the NWB format (and the many other formats that NWB supports) at: <http://nwb.slis.indiana.edu/doc.html>

Download the most recent version of Network Workbench at: <http://nwb.slis.indiana.edu/download.html>

Working With Downloads from Thomson Web of Science in Network Workbench

For this example, we will use a combination of files in Thomson's tagged field format (ISI format). The combined file contains the papers authored by neuroscience researchers at IU and IUPUI, and was created by concatenating files downloaded individually for each researcher.

- 1.** Go to File -> Load and Clean ISI File and select all.isi. This loads the specialized ISI format into a table, then eliminates rows with duplicates of the unique ID field Thomson uses. These duplicates exist because we have combined the files of multiple researchers, and many of them have coauthored papers.
- 2.** Select the second entry in the Data Manager on the right, which probably reads "3889 Unique ISI Records." Some of these records are for papers with very large numbers of authors, one even approaching 2,500. These papers are less interesting, so we will remove them.
- 3.** Under Preprocessing, choose Filter By Number. The default arguments for this will remove rows that have over 100 entries in the AU (author in ISI format) column, separated by a pipe character. This is what we want, so click OK.
- 4.** Choose the newly created piece of data in the Data Manager. Select Extract Co-Authorship Network, under Preprocessing. This will generate two files—a network and an author file. The latter can be used to merge authors that are incorrectly represented as two nodes in the network, but we will not use it in this tutorial.
- 5.** Select the generated network file. This network has a very large number of edges, so for the purposes of visualization we will sample from it. Go to the Preprocessing menu again and choose Edge Sampling. For the number of edges to sample, enter 7,500.
- 6.** Select the sampled network. After sampling, there will be many small components. The interesting part is usually in the largest component, so go to Analysis -> Unweighted & Undirected -> Weak Component Clustering. For the number of clusters, put a small number like three or five. This will create several new networks from the largest components in the parent network.
- 7.** Choose the component with the most nodes and select GUESS from the Visualization menu. If you are on Windows, after waiting a little while, quit Network Workbench. There is a bug that prevents GUESS from fully starting until NWB quits.
- 8.** After GUESS loads, choose Run Script... from its File menu. There, select the neurovis.py file. This will perform steps to create a nice visualization from datasets like the one you have generated.