Graph and Web mining – Datasets for Project

There are two types of data on which you can test your implemented project. There is data of synthetic graphs prepared by Laura Langohr, and real data links.

The synthetic data is very simple and is composed of a set of nodes in the format of (node-id label) and a set of edges immediately afterwards (only a blank line between both parts) in the format of (node-id1 node-id2).

Real data is described in the provided links. For real data you may also use your own data provided you give a full description of it and put it on the course site if its not confidential. If its confidential just give a detailed description.

Synthetic data

There are two types:

1) Two single graphs datasets one of 400 nodes – http://www.cs.helsinki.fi/u/langohr/graphmining/assignments/single-large-graph_400-nodes.txt,
   the other of 2000 nodes – http://www.cs.helsinki.fi/u/langohr/graphmining/assignments/single-large-graph_2000-nodes.txt.
2) 600 small size graph transaction files (up to 100 nodes), to be used in the transaction setting case. They are located in http://www.cs.helsinki.fi/u/langohr/graphmining/assignments/transaction-graphs.zip

Real data

You may use the following links:

1) Portion of the Movielens database, a database on movies and their raters and rating - http://www.cs.helsinki.fi/u/langohr/graphmining/assignments/movielens.zip
2) Portion of the IMDB database, a database on movies, actors, directors, etc in an XML format - http://www.cs.helsinki.fi/u/langohr/graphmining/assignments/imdb.xml
3) Links to chemical databases - http://www.cs.ucsb.edu/~xyan/dataset.htm and http://www.predictive-toxicology.org/ptc/#InitData
4) Link to social network data - http://snap.stanford.edu/data/index.html

Some of these databases are very large so you don't have to use all of the data.