Frontiers of Network Science 6250/4250, Fall 2018
Assignment 1, due at 11:59 pm on Monday, October 29th

Select three different real world networks for analysis. One network can be from your own project or research. Two other networks (or all three if you don't have your own network) should be taken from one of the following collections:

- SNAP repository (http://snap.stanford.edu/)
- Koblenz set (KS) (http://konect.uni-koblenz.de/)
- Network Repository (NR) (http://www.networkrepository.com/)
- Pajek datasets (http://vlado.fmf.uni-lj.si/pub/networks/data/)
- Mark Newman's Collection (http://www-personal.umich.edu/~mejn/netdata/)
- DIMACS Challenge Graphs (http://www.dis.uniroma1.it/challenge9/download.shtml)
- UF Sparse Matrix Collection (http://www.cise.ufl.edu/research/sparse/matrices/index.html)
- Laboratory for Web Algorithmics (http://law.di.unimi.it/datasets.php)

The network has to have at least 500 nodes and the number of edges has to be at least 2,000. We encourage you to select the networks that your computer is capable of processing, given your choice of tools, so no larger than 2,000 nodes and 10,000 edges.

Send your list of three networks for approval (email: szymab@rpi.edu), specifying where you obtained the networks (please, provide an URL) and its size to ensure that all students work on different networks. If it turns out that the network you wish to analyze has already been taken by someone else, we will ask you to select another network. So, the earlier you send us your selection, the better are your chances of getting the networks you want. Do not start working on the networks until you receive an OK from us.

For each real network, create two artificial networks having the same number of nodes and similar number of edges. The first artificial network should be ER random graph, while the second BA scale-free network. For ER graph, compute the average degree of the nodes in your real network to get similar number of edges in the generated random graph. For BA network, chose the minimum degree to get the number of edges in scale-free network closest to the real network. So you will work with 3 real networks, 3 ER networks and 3 BA networks, where each of the group of networks (one real and its corresponding ER graph and BA network) have the same number of nodes.

For each of the networks:

1. Provide a detailed description of a network. Don't forget to specify the number of nodes and edges, what they represent, how, when and by whom the network was collected, what its significance and meaning are, whether the edges are directed and weighted, etc.

2. Compute the global properties of a network: (i) the diameter, (ii) the number of connected components (for directed graphs, weakly and strongly connected) and the size of the largest one, (iii) the node average degree.

3. Compute and plot: (iv) the degree distribution (for directed graphs, also in-degree and out-degree distributions), (v) the length distribution of the shortest paths, (vi) the clustering coefficient distribution, (vii) the betweenness centrality distribution, (viii) the connected components size distribution.

   Compute and list also (ix) the average values, and (x) variances for each of the measurements (iv)-(viii).

   Make sure that when needed, you use log scale on vertical axis or log on both axes (log-log plot)
   (*) For each distribution, on the same plot also plot the best fit line.

4. For each real network, identify the most "important" nodes in a network, using different ways to approach an importance.

   For real world networks, discuss how the nodes you identified as the most important correspond to the semantics of nodes relative to the domain which nodes in your networks represent. (E.g., in our German boys' school class network, Lasch has a high degree but low ranking because he was "buying into" popularity with sweets and money from his grandmother.)

5. Compare your networks using the measurements made in tasks 2 through 4, and discuss differences between the metrics and then decide what type of the real network you are dealing with (for example, random (ER), scale-free (SF), or other (like regular grid, bipartite network or undefined). Justify your answers.

6. Render a high quality (vector) graphic representation of each network and include it in your report. If the network is too large to have all nodes and edges drawn directly without affecting the presentation quality of the figure, you may use filtering, collapsing expansion, hierarchical representation, or other techniques to reduce visual clutter.

   Optional: Make the visualization of your network visually appealing (i.e., legible, with proper layout and labels when necessary) and meaningful (provide visual cues that should help readers understand your analysis of the network, e.g., different colors and sizes of nodes reflect different values of metrics which you discuss in your analysis). Provide the legend explaining the meaning of different colors, sizes, types of lines, etc. and justify your choices.

You can use whatever tools (either your own or third party) you deem appropriate for the job. If you are using a third-party tool (e.g., Gephi, Neo4j, etc.), it has to be available free of charge (or at least a free of charge fully functional evaluation version should be available). Please document which tools you use for which task, including the URL of the tool Web site.

If you are using your own tools (e.g., you are writing your own programs), please specify which ones and for which tasks and provide the source code and executable for ThinkPad running Microsoft Windows 10 along with the relevant instructions on how to run them.

If you are using third-party frameworks or applications (e.g., Matlab, Microsoft, Excel, etc.) indicate which tool was used for which task and provide all user files (".m", ".xls", etc.).

In your report, provide a brief justification for your choice of a particular tool for a particular task.

The maximum numbers of points assigned to Tasks are as follows:
Task 1: 5 pts, Task 2: 5 pts, Task 3: 10 pts, Task 4: 10 pts, Task 5: 10 pts, and Task 6: 10 pts.
Partial answers of partially correct answers will earn partial credit.
Optional task: 10 points.

The assignment can be done by each student individually or by a two person team. Each team member needs to email me no later than the end of Thursday, October 11 the team members. No collaboration is allowed outside a team or between students solving the assignment individually.

You can use whatever tools (either your own or third party) you deem appropriate for the job. If you are using a third-party tool (e.g., Gephi, Neo4j, etc.), it has to be available free of charge (or at least a free of charge fully functional evaluation version should be available). Please document which tools you use for which task, including the URL of the tool Web site.

If you are using your own tools (e.g., you are writing your own programs), please specify which ones and for which tasks and provide the source code and executable for ThinkPad running Microsoft Windows 10 along with the relevant instructions on how to run them.

If you are using third-party frameworks or applications (e.g., Matlab, Microsoft, Excel, etc.) indicate which tool was used for which task and provide all user files (“.m”, “.xls”, etc.).

In your report, provide a brief justification for your choice of a particular tool for a particular task.