Frontiers of Network Science 6250/4250, Fall 2023
Assignment 1, due at 11:59 am on Monday, October 16th

Select two different real-world networks for analysis. One network can be from your own repository. Another network (or both if you do not have access to some repository) should be taken from one of the following repositories:

- 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)
- Colorado Index of Complex Networks (ICON) - Large collection of networks described and indexed by research group of Aaron Clauset.
- Network Datasets by Eric D. Kolaczyk.
- Linton Freeman's Network Data - Over 300 datasets of all sorts, in UCINET format.
- Network Science Book - Network Datasets - Network data sets from Albert-László *Network Science* Barabási book. Includes data on IMDB actors, arXiv scientific collaboration, network of routers, the US power grid, protein-protein interactions, cell phone users, citation networks, metabolic reactions, e-mail networks.
- Nexus - Repository of network datasets in GraphML and igraph formats.
- Pajek Datasets.
- Siena Datasets.
- UCI Network Data Repository.

The network must have at least 500 nodes and the number of edges must 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 10,000 nodes and 30,000 edges. You can also use any network from a repository not listed here but you need to provide me with the link to the source, and select networks with the sizes described above.,

Send your list of two networks for approval (email: boleslaw.szymanski@gmail.com), specifying where you obtained the networks (please, provide an URL) and their sizes to ensure that all students work on different networks. If it turns out that another student already chose any of the networks you wish to analyze, we will ask you to select another network. Hence, 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. The deadline for choosing the networks is before Lecture 11 on October 2nd,

For each of the two real networks, create two artificial networks having the same number of nodes and similar number of edges. The first artificial network should be an ER random graph, while the second should be a BA scale-free network. For ER graphs, compute the average degree of the nodes in your real network to get similar number of edges in the generated random graph. For each BA network, chose the minimum degree to get the number of edges in the scale-free network closest to the real network. Thus, you will work with two sets of three networks: real , its ER version and its BA version, with all three networks having the same number of nodes and similar number of edges, but at least two different degree distributions.

For each of the six networks:

1. Provide a detailed description of two networks that you chose. Do not 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, radius, and average eccentricity, (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 and degree variance.

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

   Compute and list (viii) the average values, and (ix) variances for the measurements (v)-(vii).

   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.

   Hint: for 2(i) and 3(v) repeatedly use breadth first walk from the node to finds its longest shortest path (one path per node)

Optional: Render a high quality (vector) graphic representation of each of your two chosen networks 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. 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. 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), it must 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) indicate which tool you used and 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: 10 pts, **Task 2:** 15 pts, Task 3: 15 pts. **Optional task:** 10 points
Partial answers of partially correct answers will earn partial credit.

Each student should do the assignment individually; students should not collaborate on it.
Submit your solution to email of TA, with necessary files with answers and plots as attachments.