Complex network analysis project

In this sheet we describe the complex network analysis project you have to fulfill in order to complete the exam. This project is **individual** and it concurs to the 30% of the final mark.
Within this academic year, there is no deadline to hand the project in, but the exam will be registered only after both the written exam and project marks have been assigned.

To hand the project in, write an email from your institutional address to lorenzo.dallamico@unito.it **and** michele.tizzani@unito.it . The object should be "YOUR NAME - COMPLEX NETWORK PROJECT". Attach to the email the **pdf** file of your report together with the commented codes used to produce the results.

Project description

You are asked to download a network of your choice and to perform a series of analysis on it, based on the topics discussed during the course lectures. You will have to write a report of your findings describing in detail the results you obtained. This report should include a description of the data you used, the results of the simulations you run and an extensive discussion thereof. For all assignments we require the presence of three sections: i) *Network metrics* in which you describe in words what the graph represents and you run basic analysis on it (see below); ii) *Community structure* in which you analyze the class structure of the graph; iii) *Epidemic simulation* in which you choose an epidemic simulation on your network to choose between SI, SIS and SIR.

Data

Select one of the datasets from the link: http://snap.stanford.edu/data/index.html. As you can see, in this repository, graphs are already subdivided into groups. For all of them you have the number of nodes and edges forming the graph. Some of them have millions of nodes and running your code can be particularly challenging. For this reason we suggest to consider graphs with at most 50K nodes or to consider only a subgraph of smaller dimension. If you opt for the latter option, specify it in your report, explaining the how and the why of your sampling strategy.

We now describe the content of the sections of your report.

Network metrics

In this section you are asked to describe the graph you are dealing with. Some of these data can be obtained on the website and should be compared to check the consistency of your findings. You should have a code capable of computing the following quantities:

- Graph size: number of nodes
- Number of edges
- Node degree and degree distribution
- Connected components (choose whether to work with the whole graph or only with one connected component). Describe how many they are, how big they are, the average degree within each connected component.
- Centrality of each node (choose any measure of centrality you like, except the obvious degree centrality).
- For each node find the longest shortest path and find the graph diameter
- Clustering coefficient of every node.

All these quantities should be reported or plotted and commented. For instance, you could try to answer to questions of the type "*Is the network small world?", "How does the diameter compare to the average longest shortest path?".* At the end of this section, you should have given a comprehensive presentation of the graph you are working with and have a clear picture of the graph properties and possibly relate it to the type of data it represents.

Community detection

In this part of the project, you should perform community detection on your graph to unveil its class structure (if any). Does the graph have communities? How many? Is there a hierarchical structure? To answer this type of questions, use any community detection algorithm you like from the set of those presented during the course. It is crucial however that you are capable of commenting and interpreting your results, taking the weaknesses of each algorithm into account. A useful approach is to compare different algorithms and in the end make a global sense of the mesoscale structure of the graph.

Epidemic simulation

Once you thoroughly analyzed the graph, you should run a simulation on it. Perform the following tasks:

- Simulate one of the three epidemic models studied in the course (SI, SIS, and SIR). Run the simulation multiple times initializing the epidemics from a random node of the network at each run.
- Plot the epidemic threshold as a function of the model parameters and the number of infected individuals over time.
- What is the effect of the most central node in the spreading of the disease?
- Suppose you find K communities in your network. Compare the results starting the epidemic from K nodes in one community with what happens starting the epidemic from K nodes but one for each community.

The output of your project should be a report of all these results, accompanied by the documented codes used to obtain the results.