

# Project 1: Analysis of a Network

## COMS 6998-006 Network Theory, Spring 2008

Network Due: February 18, 2008  
Entire Project Due: February 25, 2008

### 1 Overview

The first assignment is to create a network by gathering data, analyzing it using tools such as clairlib and pajek, and documenting your results in a 6-page Physics Review (PR) Style paper.

### 2 Italian Network

[10 points]

In order to familiarize yourself with the network tools, the first portion of the project is to analyze the Italian Network which is located at [http://www1.cs.columbia.edu/~coms6998/pajek\\_example.net](http://www1.cs.columbia.edu/~coms6998/pajek_example.net). You are required to:

1. Run clairlib on the Italian network.
2. Run pajek on the Italian network.
3. Report results, including images and tables.

### 3 Data Collection

[Total: 10 points + 10 Extra Credit]

You can choose to build your own network or use an existing network. It is strongly preferred that you build your own network. [10 extra points]

#### 3.1 Build Your own Network

[10 points + 10 Extra Credit]

The main goal of this assignment is to build your own network. The networks can be text or web-related as well as social, biological, etc...

1. Gather data. This may include writing some of your own code.
2. Explain how you built your network.

### 3.2 Use an existing Network

[10 points]

If you choose to use an existing network it is preferred that you analyze one that has not been described in a paper yet.

1. Choose existing data.
2. Explain how the data was generated by the owner (If you can't find how the data was generated, explain how you would have generated the data).

## 4 Analysis

[10 points + 10 Extra Credit]

1. Analyze your network using your favorite tools (see resources) or your own code [10 points, up to 10 extra points for your own code].
2. Write a detailed network analysis in your report including images and tables. Mention what tools you used to generate the data.

## 5 Report

[Total: 60 points]

You are required to write a 6 page PR style report documenting your work and results received from the Italian Network, Data Collection, and Analysis. Below is a list of some results that should be included in your report.

1. What data did you choose and how did you get it? Why is it important? [10 points]
2. What type of distribution does your data appear to have? Show graphs to illustrate. [10 points]
3. How many nodes and edges are in your network? what is the diameter? Average Degree? etc... [5 points]
4. What are the connected components? What is the clustering coefficient? Shortest Paths? Triangles? etc... [5 points]
5. Display visualizations of your network. If your network is too large, show samplings. [10 points]
6. What sort of interesting properties does your network contain? Did you come across anything unusual or unexpected? [10 points]
7. How does your network compare to a similar random graph? [10 points]

## 6 Choosing A Network

**Due: February 18, 2008**

You are required to have chosen a network by Monday, February 18th. Send an email stating the following to Drago (radev@umich.edu) and Sara (ss3067@columbia.edu):

- Your name.

- Description of the network (2 paragraphs).
- Size (n, m), if available otherwise give an estimate.
- How it was/will be collected.
- List five sample edges.

For example:

- Name: John Doe
- Description: computational linguistics program committee data set. Two people are connected if they were on a program committee (or senior program committee) together. The conferences include ACL, NAACL, HLT, EMNLP, EACL, and COLING and all their associated workshops from 1996 to 2007. One sample workshop is listed here: <http://www1.cs.columbia.edu/nlp/acl05soft/>
- Size: There are 150 program committees of 10-60 people each. Each person is therefore associated with a large number of neighbors. Estimate for  $n = 3,000$  and  $m = 100$ .
- Collecting Data: by searching the web for lists of PCs and manually building the network
- Examples of edges:
  - Tilman Becker <-> Nizar Habash
  - Brian Roark <-> Nizar Habash
  - Tilman Becker <-> Brian Roark

## 7 Resources

- A list of possible network ideas is listed at <http://www1.cs.columbia.edu/~coms6998/datasets.htm>.
- The Italian network can be downloaded at [http://www1.cs.columbia.edu/~coms6998/pajek\\_example.net](http://www1.cs.columbia.edu/~coms6998/pajek_example.net).
- A tutorial on how to run clairlib for this assignment is located at [http://www1.cs.columbia.edu/~coms6998/Clairlib/clairlib\\_network\\_analysis.html](http://www1.cs.columbia.edu/~coms6998/Clairlib/clairlib_network_analysis.html).
- Software available on the clic machines, as well as instructions to run them, is listed at <http://www1.cs.columbia.edu/~coms6998/software.html> (Software: Pajek, Jung, Guess, NetworkX, Clairlib).
- Instructions to generate a PR Style paper are describe at [http://www1.cs.columbia.edu/~coms6998/pr\\_style/instructions.html](http://www1.cs.columbia.edu/~coms6998/pr_style/instructions.html).

## 8 Submission

Your submission should include your dataset, any code you wrote, and your report.

You need to submit a TAR file with the following files:

- The raw data used to build the network.
- The network in Pajek and/or clairlib format.
- The PR-style paper (including latex + pdf + any images in separate files + makefile).
- Any software needed to preprocess or analyze the data.

Please name your files and directories using your last name and the homework number, e.g., John Smith will name his files and directories Smith1.

```
Smith1.tar:  
Smith1/paper  
Smith1/paper/Smith1.tex  
Smith1/paper/Smith1.pdf  
Smith1/data/Smith1.net  
Smith1/data/...  
Smith1/code/...  
Smith1/...
```

## 9 Grading

- 10 points for analysis of Italian Network
- 10 points for gathering data (10 additional points for new data)
- 10 points for running analysis tools (up to 10 additional points for your own code (only if the code was necessary))
- 60 points for report:
  - What data did you choose and how did you get it? Why is it important? [10 points]
  - What type of distribution does your data appear to have? Show graphs to illustrate. [10 points]
  - How many nodes and edges are in your network? what is the diameter? Average Degree? etc... [5 points]
  - What are the connected components? What is the clustering coefficient? Shortest Paths? Triangles? etc... [5 points]
  - Display visualizations of your network. If your network is too large, show samplings. [10 points]
  - What sort of interesting properties does your network contain? Did you come across anything unusual or unexpected? [10 points]
  - How does your network compare to a similar random graph? [10 points]

**Total:** 90 points + a possible 20 Extra Credit points = 110 points.