

WMR 2014 Final Term


May 19, 2014

Final Assignment

The Final Term consists on the analysis of real world complex networks. Each group¹ has to:

- i) select and analyze **two** networks (avoiding overlap with other groups);
- ii) solve **two** of the proposed exercises and discuss results and methodologies in a written report.

Surveys have to be sent to pedre@di.unipi.it, giulio.rossetti@isti.cnr.it, lpapalardo@di.unipi.it in pdf format (using as subject [WMR2014 Final Term]). The Final Term projects will be discussed during an oral dissertation.

 **WARNING:** Students who did not participate at the translation project must solve an extra exercise.

Oral Exam: Agree a date for the oral dissertation with the professor.

Submission: the project must be submitted within the summer exam session (at least 3 days before the date of the discussion).

¹Composed at most of 3 students.

Exercises

[Mandatory]

– **Network analysis:**

Characterize the chosen networks through the analysis of all the basic measures introduced during the course. Compare obtained results and highlight the identified differences.

[To choose]

– **Link Prediction**

Partition each network in a training (80% of the edges) and a test set (20% of the edges) and apply some of the classical unsupervised link prediction approaches introduced in [1] (i.e. Common Neighbors, Adamic Adar, Jaccard, Preferential Attachment). Discuss the prediction accuracy as done in the referenced paper.

– **Community Discovery**

Apply on each network at least two community detection algorithms: K-Cliques, Girvan-Newman [2] and DEMON [3]. Discuss the obtained results (distribution of nodes in communities, number of communities, Community sizes, average density, etc. etc.).

– **Diffusion**

Implement and test the SIR, SIS or SIRS models explained in [4] (chapter 21: Epidemics). Alternatively apply a cascade model (as introduced in [4] chapter 19: Cascading behaviors in networks).

– **Ranking**

Apply the PageRank and Hub & Authority ranking algorithms to the selected networks: compare and discuss the obtained results.

Datasets

Several network dataset can be gathered, among all the possible online sources, from the following repository:

- <http://snap.stanford.edu/data/>
- <http://wiki.gephi.org/index.php/Datasets>
- <http://www-personal.umich.edu/~mejn/netdata/>
- <http://www.giuliorossetti.net/about/en/ongoing-works/datasets/>

⚠ Tips:

Select networks with at most 10-20k nodes in order to avoid computational issues.

⚠ Tips:

Prefer networks described in edgelist format (an edge per row). An example is:

```
nodeid nodeid
0 1
1 2
2 0
...
```

Code and libraries

To solve the mandatory exercise you can make use of Cytoscape² and/or Gephi³. To approach the remaining exercise you will need to write some code: you can use any programming language and library you prefer. However, especially for those who are not familiar with code developing, we suggest to use the networkx⁴ Python library due to its simplicity⁵.

Implementations of the DEMON algorithm and of classic link prediction approaches can be downloaded from:
<http://www.giuliorossetti.net/about/ongoing-works/material/>.

²<http://www.cytoscape.org/>

³<https://gephi.org/>

⁴website: <http://networkx.github.io/>

⁵A suggested Python oriented IDE is Pycharm: <http://www.jetbrains.com/pycharm/>

References

- [1] David Liben-Nowell, Jon M. Kleinberg: The link prediction problem for social networks. CIKM 2003: 556-559
- [2] Girvan M. and Newman M. E. J.: Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 78217826 (2002)
- [3] Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi: DEMON: a local-first discovery method for overlapping communities. KDD 2012:615-623
- [4] David A. Easley, Jon M. Kleinberg: Networks, Crowds, and Markets - Reasoning About a Highly Connected World. Cambridge University Press 2010