## Social Network Analysis, Dr. Pierpaolo Dondio

## Module Research Project (50% of final exam mark)

The aim of the project is to deliver a paper-like document describing a research in Network Science. The document could be an experimental project, simulation or a data-driven study.
The project can be delivered individually or in groups of 2 or 3 students.
**Deadline: Wednesday 16th December 2020 end of the day.**

The document has to be formatted as an academic paper, including:

- Abstract
- Introduction describing the aim of the research and its research question(s)
- Related Works in the area
- Methodology, including:
    - Data gathering, data preprocessing and description of the dataset (if applicable)
    - Research Methodology used
- Experimentation and Results
- Discussion
- Conclusions

The paper will be in the region 4500-6000 words. Different kind of projects are possible, for a list of suggestions and project ideas see later in this document.

**Marking Scheme**

In a conference-like peer-review process, your "paper" will be reviewed and scored.

All submissions will be peer reviewed and evaluated on the basis of the overall quality of their technical contribution, including criteria such as originality, soundness, relevance, significance, quality of the written presentation, quality of the oral presentation, methodology, dataset gathering effort, implementation and understanding of the state of the art. In details the criteria will be:

1. Originality of work: 5 marks
2. Potential Impact of results: 3 marks
3. Quality of the methodology: 7 marks
4. Quality of execution: 5 marks
5. Complexity of the project:10 marks
6. Quality of presentation: 5 marks
7. Adequacy of citations: 5 marks
8. Quality of Oral presentation and defence: 10 marks

Each of them will be equally important and the total sum will be 50 marks.

## Some Project Ideas

### 1. Cultural Network in Wikipedia

The aim of the project is to analyse the relations between cultures in Wikipedia by building a network(s) where nodes represents one of the 250+ Wikipedia versions and there is a link between two nodes if the two Wikipedia versions have something in common (for instance, they share the same articles or they assign to a group of articles the same importance… (see the references and papers loaded on webcourses for some ideas).

The idea of the project is to analyse such network of "cultures" in order to identify cultural similarities, recurrent patterns, cluster of cultures, dominant cultures, isolated cultures and explain/investigate the factors that could have generated such network, such as international diffusion of a Language, colonialism, geographical location, common beliefs/religion between countries, former political union, immigration.

### 2. Disease Diffusion over the Airport Network

This is a simulation project in the same line as the paper by Lada Adamic (on webcourses).
The idea is to simulate the spread of a disease over the airport network. It is a simulation approach where, according to a probabilistic model, an infection can start in one of the airports and with another probability it can spread to connecting airports. Each airport has a number of *medical resource units* to fight the spread of the disease or to prevent it.
Resources and disease probabilities have to be assigned based on some data/assumptions, considering for instance the expenditure of each country in Health. Each airport has a set of resources to stop the diffusion and resources can be allocated dynamically. The simulation wants to find the most vulnerable airports, find to what extent a disease can spread and find the best strategy to allocate resources to minimize the diffusion. Various approach can be tried and tested by varying the disease probability (more or less contagious, incubation time, duration). Results can be compared to the following baseline strategies:

- A random allocation of resources
- A strategy that allocates resources to airports based on airports position in the network (measured by centrality for instance)

See Lada Adamic's paper on Brightspace for a good example of methodology and evaluation. The idea is to replicate her methodology
In general, this topic is about simulations over a network. Anything rather than diseases – such as traffic, ideas, power grid, movements of people – can be simulated…

### 3. Documents classification using Network Modularity and Communities detection algorithms

Automatic classification of documents is a classical text-mining application. Usually, documents are preprocessed into a document vector representation, where each document is represented by a set of terms used in the document, with a weight attached, measuring the importance of each term for the document. Common measures for the weight of each term is the frequency, or the TF-IDF metrics. Based on this vector-based representation, documents are then classified using centroids and cosine similarity, or using Support vector machines classifiers. A good introduction is here:

*Han, Eui-Hong Sam, and George Karypis. "Centroid-based document classification: Analysis and experimental results." European conference on principles of data mining and knowledge discovery. Springer Berlin, 2000.*

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.608.460&rep=rep1&type=pdf

The aim of this project is to investigate the feasibility of applying network partitioning algorithms (such Newmann's modularity, hierarchical clustering, egde-betweeness clustering) to documents classification and compare the accuracy of these predictions with classical approaches, or mixing both of them in an ensemble classifier.

The idea is to create an undirected network where nodes are documents and there is an undirected link (probably weighted) if the two documents are *"close enough"*.

Lada Adamic's paper on Network Ingredients can give some ideas about building networks between words or group of words. The pointwise mutual information could be used as a metrics, or the same starting representation (vectors of TF-IDF and cosine similarity) could be used. However, now the network clustering algorithms are used to group documents instead of K-means, centroid distance or SVM.

Different portion of documents will be used for training/testing/validating the algorithm and the usual metrics (accuracy, precision…) will be used to compare network-based and distance-based classifiers.

## 4. Social clusters at DIT and beyond

This is a project with a strong sociological aspect. The aim is to study social clusters and relations among individuals (DIT students could be a possible group).
The idea is to design an anonymous survey where participants are asked to answer questions regarding what they like/dislike, asked to rank items and so forth… (the content of the survey depends on the aim of the study). They are also asked to provide some demographic information such as gender, age or other factors defining social groups (nationality for instance..).
Using the data collected during the survey, a bi-partite graph can be drawn connecting individuals to answers. From that graph an undirected (potentially weighted) graph among individuals can be built. Then, there could be two different directions.

A first, more sociological analysis, will focus on describing and analysing the network, and try to justify the formation (or absence) of clusters, and it will focus on testing small worlds hypothesis, analysis of motifs, presence of weak ties, analysis of the most influential nodes, network resilience and so forth. The analysis of the network could reveal interesting patterns or help answer some hypothesis on social groups.

A second approach is closer to a predictive analytics study, and it tries to predict the features of the node (=the demographic information collected) as target variables. Multiple predictive models can be tested using as features both the answers (as in a classical data mining model) or the network features of each node (node belonging to a group, its centrality measures, its position, its distance to other nodes..)
It will be interesting to check if the addition of network features can improve the quality of the predictions.

### 5. Analysis of SO tag questions

This is a big-data Network visualization and analysis project.

The aim of the project is to visualize and analyse the network of tags of the entire Stack Overflow dataset. Each SO question has a set of tags associated with it. For the scope of this project, those tags are enough. However, you are free to do a more detailed study and automatically tag questions (or a sample of questions) with a more fine-grained set of tags, maybe mined from the text of the question/answer.

The resulting network of tags represents a network of computer science concepts and topics, with their complex relationships. It is a very large network and visualize it in a convenient way is not an easy task. It might require clustering, partitioning and filtering. The first task of the project is to find software and ways to present the network.

Regarding the analysis, the network could be used to analyse the relationship among computer science/programming concepts. By using generalization/specification it will be possible to produce a list of pre-requisites necessary to understand a topic (represented by the tag), or the network could be used by learners/teachers to identify key concepts versus peripheral concepts, concepts linking different topics (and therefore fundamental) and so on.

For instance, in order to connect to a MYSQL database via php and execute a join query we need to understand the concept of foreign key, table, inner join, some SQL query syntax, some PHP, the concept of cursor and so on.

A complete dump of the Stack Overflow dataset can be downloaded here (https://archive.org/download/stackexchange).

**Other ideas**

6. Network of Literary documents, or music lyrics in order to classify songs/poems, cluster the network, identify influential artists, predict the success of a song/book…
7. Resilience of a network, or how a network can be made more efficient by changing its topology..

LOOK AT THE PAST PROJECT SAMPLES ON BRIGHTSPACE AS WELL

## Your own ideas are also appreciated!