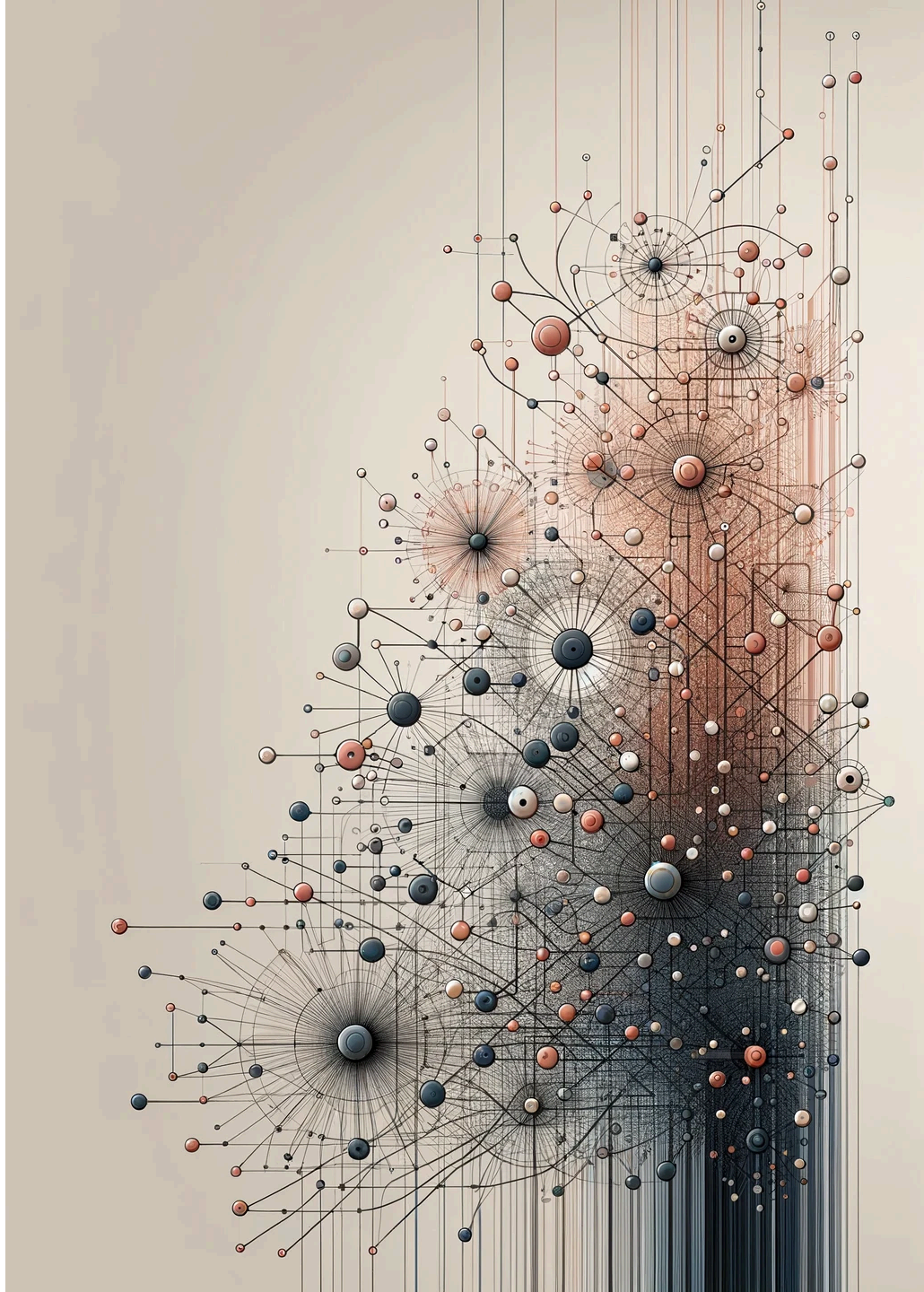
UNIVERSITÀ DI TORINO

Analisi e Visualizzazione delle Reti Complesse

Final Project

Prof. Rossano Schifanella



UNIVERSITÀ DI TORINO

Scope

- The objective of the final project is to let the students play with a **real-world dataset** applying the **theoretical concepts** and the **practical tools** we have introduced in class.
  - and perhaps exploring something new in autonomy!
- After selecting a topic and datasets, the goal is to present a **narrative through quantitative exploration**.
- The dataset should allow the formalization of the problem using **networks**.
  - networks have to be present, however, other types such as geographical data may also be included.
  - when applicable, it is recommended to explore heterogeneous contexts.
- **[9 credits]** The narrative should contain a **visual exploration** of the domain and the results should be represented in a **data visualization**.
  - **interactive** vs **static**
  - **same domain** vs **different domain**

Organization

- The project can be developed in **groups of maximum 3 students**.
- You have to **register** your project at **this link** to specify the group and the topic.
  - **deadline: the end of the course.**
- Groups and topics *should not* be changed after the registration.
  - of course, this could happen for many unpredicable reasons, so in this case, please, reach out and we will find an alternative solution.
- Groups and topics will be presented in the last class of the course. You will be asked to:
  - Prepare a short presentation (5min-8min) explaining the idea and the datasets you intend to use
  - Present to the class your project
  - Receive feedbacks
  - more info will be shared towards the end of the course

UNIVERSITÀ DI TORINO

What to submit

- A **report** describing the project and its contributions. Generally, it should contain the following sections:
  - Introduction and contextualization
    - What is your project about and why it is worth exploring it
  - Methods
    - What methods did you implement in the analys and visualization pipelines
  - Results
    - What are the main results and observations you extracted from your analysis
  - Discussion
    - Elaborate on the possible implications of your observations, or why you made some methodological choices over other available alternatives, explore alternative scenarios
  - Conclusions and extensions
    - What did you learn and how this could be extended if you have time?
- This organization is arbitrary and serves as a template to simplify your work.
  - **You are free to use your own structure!**
- We expect **concise** yet **complete** reports.

UNIVERSITÀ DI TORINO

What to submit

- The **code** implementing:
  - (a) the network analysis
  - (b) the visualization
- Depending of the type of project it be in the form of:
  - a series of Jupyter notebooks
  - an interactive web app
  - an Observable notebook
  - combinations of them
  - others
- All the material should be packed in an archive and submitted through Moodle **at least 3 days before the exam session**.
- The link will be active at the end of the course.

General rule (probably the most important!)

If you have **any doubt** reach out and we can discuss alternatives that **fit best your case study and project**.

Exam

- The exam is organized in two main parts:
  - **project discussion (40%)**
  - **oral examination (60%)**
  - and remember the 10% bonus for the active participation in class
- **Project discussion:**
  - the group/student presents the project usually with the help of a presentation
    - the presentation is not mandatory, however, it usually usueful to organize to discourse
  - the goal is to show what has been done, the results, the methodology, run a demo if you created an interactive data viz and so on.
  - questions on anything related to the project, the submitted report, the code, will be asked.
  - **all the member of a group must discuss the project in the same session**.
    - Only in specific, motivated cases we can break this rule.

Exam

- **Oral:**
  - the oral part of the exam will test your knowledge of **all the concepts introduced in course**.
  - everything we discussed that has not be marked as optional is a potential candidate for a question.
  - it is possible that a question involves the solution of a **simple exercise** involving theory constructs.
    - a theory question could be practical, e.g., given this networks, compute the betweenness centrality of the nodes
  - each student in a group will be evaluated separately
    - you should expect around 3 questions per candidate
    - this is just an **estimate**, there will be variability

UNIVERSITÀ DI TORINO

What you can focus on?

- Generic description of the network. Calculate structural measures and plot them whenever it is possible/significant:
  - Distances: average, distribution
  - Degree: average, variance/standard deviation, degree distribution (some fit? Does it follow a power law? If yes, is it in the scale-free regime?)
  - Clustering coefficient
  - Largest connected component size
  - Degree correlation: neutral, assortative, disassortative?
  - Are there communities? Can you properly show them with an appropriate layout? Can you discuss them?
  - Centralities
  - Can you analyze homophily?
- Try to interpret the results of these measures, and comment/discuss results.

What can you focus on?

- Dynamics:
- Once you have studied your network, and you have a general understanding of its structure, you can use it to simulate some dynamic processes, e.g.:
  - Behavioral cascades
  - Diffusion of innovations
  - Epidemics
  - ...
- What could you expect to happen over that network when some of these models is simulated?

UNIVERSITÀ DI TORINO

What can you focus on?

- Generative models:
  - You might need to create your own artificial random networks for comparison purposes
  - Erdos-Renyi
  - Watts-Strogatz (small-world)
  - Configuration model (degree preserving)
  - Barabasi-Albert (preferential attachment)
  - Stochastic bloc model
- You can use different generative models, to produce comparisons, varying some parameters (e.g., linking probability, number of edges added at each step, degree distributions, and so on)
  - For some comparisons, you may need to preserve some characteristics (e.g., degree distribution). Try to rewire properly your network in order to shuffle your data.
- You can perform on such synthetic networks the analysis that has been proposed in the previous slides, to detect differences
- Try to explain different behaviors

Datasets

- Stanford Large Network Dataset Collection
- The KONECT project
- More into the data visualization side:
  - Awesome public datasets
  - Kaggle datasets
  - FiveThirtyEight Datasets
  - data.gov
  - AWS Public Data Sets
  - r/datasets
  - Google BigQuery Public Datasets

Ispiration for projects

- MIT course on Networks (2018)
- Network Science course - Ilya Makarov (2020)
- More on the data viz side but still useful - YY Ahn (2023)

UNIVERSITÀ DI TORINO

Q&A