Part II
«Portfolio-Exam»

MADS-EMDM

This is the second part of the portfolio-exam for the Data Science course MADS-EMDM (Advanced Topics of Data Mining), worth 51% of the final grade (102 points). This part of the exam is a term paper. Students are allowed to exchange ideas. However, this is **NOT a teamwork exercise**. Every student must derive and write up their own solutions in their own words and programming style.

Your task: Conduct an a graph analysis on a suitable real-world graph dataset of your choice (see criteria below). The result will be a Jupyter Notebook containing comprehensive experiments, conducted according to the following requirements.

Read the following (3 pages) carefully – they describe the conditions and criteria for grading!

1 Submission

All parts of the experiment should be contained in one Jupyter Notebook called `experiments.ipynb`. You may submit either simply the Jupyter Notebook or a zip file containing

- the notebook,
- an optional resources folder if you have additional resources (e.g. images), and
- a data folder if your data is not simply available online (cf Section 3).

Upload your notebook or zip to Moodle **before 23:59 o'clock (German time) January 11, 2023**. **Submissions after the deadline or via means other than Moodle will not be considered!**

2 The Task

Conduct a graph analysis project on a real-world dataset. For that purpose:

1. Describe a (real or fictitious) scenario for the experiments, explaining the project and its purpose.
2. Choose a graph dataset that you like to investigate and that is suitable for your experiments.
3. Demonstrate, describe, and interpret all necessary steps for conducting a clean, stable, reproducible analysis, including steps like an exploratory data analysis and the identification of central nodes.

4. Additionally include two of the following tasks: assignment of node-roles, community-discovery, or link-prediction.
5. Conduct reproducible experiments in which you design, train, optimize, test, and evaluate multiple methods (from the lectures or other sources) that are helpful in your chosen use case.
6. Draw conclusions and explain your interpretation for the practical use case.

3 Conditions

Dataset Use real-life data. Do NOT use artificial datasets NOR datasets from the lecture!

Data Acquisition There are two possible availability scenarios for the data: A) It is available online with proper license. In that case indicate in your notebook where the data can be downloaded. B) You (legally!) obtain a dataset and share it with me via Moodle. Such data should not come with any form of NDA or other obligations. If you are not sure about these criteria regarding the dataset you consider, please contact me before you invest too much time in the experiments. If you write your own code to acquire data (e.g. querying an API), create a **separate notebook** for that purpose only and submit everything in a zip file.

Programming Language Use Python for the submission.

Language Choose between English and German for all textual content.

Code from other Sources You may reuse all the code from this module's lectures and exercises. Copying (and adapting) from other sources is allowed in small quantities – e.g. a function from stackoverflow. Quote the respective source. WARNING: Copying code in large quantities will be treated as intent to deceive and result in a score of zero points.

4 Expectations

Your final score will be composed of 25 points for story, 10 points for presentation, and 67 points for the actual experiments. (Note that poor presentation can lead to loss of points in the other two categories as well, e.g. if its too confusing!)

4.1 Story

Think of the Jupyter Notebook you submit as a report of a (small) project which you – as a data scientist in a data science company – conduct for a customer. The report should follow a straight story and must:

1. explain what the project is about, what the goal is, and what the value of achieving this goal is for the customer.
2. explain the data (e.g. the features) and the (meaningful) plan to achieve the goal.
3. contain an analysis of relevant properties of the data.

4. demonstrate your ability to tackle the customer's task – show and explain the efforts you made, and interpret the results for the customer and emphasize the customer's gain through your work.
5. end with clear conclusions, a recommended course of action for the customer, a summary of the achievements, as well as concrete perspectives for future work.

Hint: Your notebook should contain MUCH MORE TEXT, introduction, comments, etc. than the notebooks we use to demonstrate methods in the lectures or the exercises!

4.2 Experiments

When grading your experiments, I consider technical soundness, completeness, and fit to the proposed task. I expect (among others):

1. Your experiments contain all the above requirements (see Section 2).
2. Your code is executable and yields reasonable and reliably replicable results.
3. All cells of the notebook have been executed.
4. The choices of methodology (different approaches, dataset specific choices, settings of hyperparameters, etc.) make sense and are reasonably explained.
5. You have experimented with a reasonable selection of hyperparameters (just one selection is not sufficient).
6. The experiments address the influence of random choices and use appropriate steps to mitigate them.

4.3 Presentation

When grading the presentation, I will put myself into the position of your project's customer. I expect (among others),

1. that the imports are organized,
2. that the code is documented,
3. that there is no unnecessary code, no lengthy debug output, no error messages,
4. that the dataset is explained (e.g., features, domain-specific notions, and abbreviations),
5. that results are presented in tables and customized diagrams which are referenced and **interpreted** in the text,
6. that I am guided through the different parts of the experiments and told what the purpose of upcoming code blocks will be.

Final Hint: Keep in mind that this is NOT your Master Thesis. The experiments should be comprehensive and created and conducted solely by you. However, limit your work towards solving ONE task – even though along the way you might recognize other interesting angles to follow up on.