2AMS30

Network Statistics for Data Science

Lecturer: Pim van der Hoorn
Email: w.l.f.v.d.hoorn@tue.nl
Office: MetaForum 4.071B

On-campus hours:	Monday 9:45-12:30	Luna 1.240
	Thursday 13:30-16:15	Atlas 7.320
Course website:	https://canvas.tue.nl/courses/32774/	

Prerequisites:

To effectively participate in this course, students are expected to know some basic concepts from probability theory, such as random variables, expectation, cumulative probability distribution, and be familiar with some basic probability distributions (Binomial, Poisson ect.). In addition, some active knowledge of coding in Python is beneficial. The student is expected to acquire any missing knowledge before it is needed for specific course activities.

Learning goals:

After finishing this course students should be able to:

1. Formulate the problem of community detection in network data.
2. Give two methods for solving the problem of community detection and identify pro's and con's of both methods
3. Apply both methods to specific instances of a community detection problem and interpret the results.
4. Formulate the problem of link prediction in network data.
5. Give two methods for solving the problem of link prediction and identify pro's and con's of both methods
6. Apply both methods to specific instances of a link prediction problem and interpret the results.
7. Describe the network science cycle for problems related to network data
8. Execute the network science cycle on a given problem for network data

General structure:

The course is structured in three parts. The first part takes one week and serves as a general introduction to networks. The other two parts each cover 3 weeks and are focused on one specific network data problem. The content is presented to students by a combination of interactive workshops, lectures, homework assignments and student presentations. The lectures and workshops are intertwined to provide a healthy mix of activities during the education slots. The main idea is that students get acquainted with the material via hands-on examples and then study the topic more in-

depth in a team and present their findings to the other students. Each part ends with an assignment where students get a real-life data set and are asked to apply the method and knowledge they learned to study this data. They then present and discuss their findings with the other students and lecturers.

For this year (2025-2026) the course topics are:

- Link prediction, and
- Community detection

Lectures
Lectures will be focused on conveying information necessary for students to analyze the data and reflect on the outcomes. The lecture will be using slides which will be made available on the course Canvas page.

Workshops
A significant part of this course will consist of workshops, where students work in groups to get hands-on experience with working with network data and addressing the network-based problems discussed in this course. These workshops will mainly consist of working out Jupyter Notebooks, which are available on the Canvas page of the course. During the workshop, a lecturer is present to help student groups with the exercises in the notebooks.

IMPORTANT: Students are expected to bring their laptops and have a working environment on them to work with Jupyter Notebooks. A simple way to get this is to install the Anaconda distribution (https://www.anaconda.com/download)

Presentations
Most of the assignments during this course end with a presentation where the students summarize their findings. Each student in the group is expected to take part in the presentation. These presentations also serve as a moment where students can discuss the topics with each other and the lecturer. Students are expected to actively participate in these discussions as well. Both the presentation and discussion are a key part of the assessment of the student (see **Assessment** for more details and a rubric).

Office hours
There are a few occasions on which no activities are planned. During this time students are encouraged to work on the course material and assignments. They can also come to the offices of the lecturers to ask any questions regarding the course.

Course resources:

Networks

- Coscia, Michele. "The atlas for the aspiring network scientist." arXiv preprint arXiv:2101.00863 (2021). Online pdf: https://www.networkatlas.eu

Community detection

- Fortunato, Santo, and Darko Hric. "Community detection in networks: A user guide." Physics reports 659 (2016): 1-44.
- Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.

Link prediction

- Kumar, Ajay, et al. "Link prediction techniques, applications, and performance: A survey." Physica A: Statistical Mechanics and its Applications 553 (2020): 124289.
- Lü, Linyuan, and Tao Zhou. "Link prediction in complex networks: A survey." Physica A: statistical mechanics and its applications 390.6 (2011): 1150-1170.
- Guimerà, Roger and Sales-Pardo, Marta. "Missing and spurious interactions and the reconstructions of complex networks." Proceedings of the National Academy of Sciences 106.52 (2009): 22073-22078

Assessment

The assessment for this consists of three parts:

1. A practical assignment on community detection (during the quartile) – 30% of final grade.
2. A practical assignment on link prediction (during the quartile) – 30% of final grade.
3. A final exam on theoretical concepts (during the exam period) – 40% of final grade.

Students need to attain at least a **4.0** on each assessment part to be eligible to pass the course. This should ensure all learning goals are at least partially attained, and this should discourage freeloading.

Practical assignments

For the two network problems in the course students are asked to **execute step 2—5 of the network science cycle** on a dataset that is assigned to them. Students will ideally work in **groups of 3-5**. Each group will be assigned a specific method they should use to analyze their data. No matter the assigned dataset and method, students will be asked to defend and/or attack its choice, and reflect on the results it produces. In principle, every student of the group is subjected to the same assessment. It will be based on the depth of understanding they show in the deliverables:

a. **Understanding –** Students can explain the idea behind the network problem, and assigned method to tackle the problem.
b. **Applying –** Students show that they are able to correctly execute the assigned method on a "new" network data-set.
c. **Analyzing** – Students are able to relate the output of their assigned method to the original problems and are able to experiment with the algorithm to obtain the "most fitting" results given their problem.
d. **Evaluating** – Students can reflect on the results they have obtained not only to assess whether their network problem has been solved, but also to critique or appraise their assigned method for their problem.

The deliverables of each practical assignment is two-fold:

I. <u>Presentation</u>

Students will give a short 10 minute presentation detailing their approach and findings, with a short 5 minute discussion afterwards. Each group is expected to participate in these discussions. Students groups hand in their slides on the day of the presentation.

During the presentations the students will be graded on two parts of the rubric: Scientific Discussion and Verbal Presentation.

It is advised to structure the presentation in line with the content needed for the poster. This way the feedback can be more directed to who the final products are graded.
Important: Because students are graded on two parts of the rubric during presentations, each student should present at least once during the course.

II. Poster
Students will create an A0-poster on their approach. For both assignments the poster can be handed in at the end of the quartile. The poster should be easy to read, with the relevant information easy to find and not be full of text.

Both presentation and poster should explicitly highlight the steps in the network science cycle. Both deliverables will be subjected to a rubric. The levels in the rubrics are based on the levels of understanding a—d. The rubric can be found in Table 1. This rubric is used both as formative and summative assessment. During the presentation, students will receive feedback on their performance based on the rubric. This will allow them to gauge how they are doing so that they can improve for the poster at the end of the topic.

The grade for each practical assignment is acquired by grading the poster based on the rubric, taking into account the presentation

Below is a list of items both the presentation and poster **must** contain:

- Problem definition
- Description of the dataset
- Brief explanation of the method and how it is applied
- Description of the obtained results
- Reflection on results

Final exam
The students will make a **three hour written exam in ANS Delft format** in which the cognitive learning goals are assessed as well as step I of the network science cycle. The exam is **closed book**, and the content of the exam is based on the **material of the lectures**. The final exam also serves as an individual assessment for each student.

Table 1 Rubric practical assignments.

Criterion	***Good*** *[80—100%]*	***Sufficient*** *[60—80%]*	***Insufficient*** *[40—60%]*	***Poor*** *[0—40%]*
Problem definition and method *Students are able to explain the idea behind their problem and method of choice both in mathematical and informal language. [20 pts]*	*Students give both a correct intuitive and a correct mathematical definition of the problem and the method to tackle it. They are able to relate the two definitions to one another.*	*Students give a correct intuitive definition of the problem. They can also formulate parts of this definition in mathematical language. Some attempt to bridge the gap between the intuitive and formal definition is made, but it does not always work.*	*Students give a definition (albeit intuitive or mathematical) of the problem. However, this definition is not fully correct.*	*Students either fail to give a definition of the problem, or they give a definition that contains a major flaw. It cannot be expected that results can be generated based on this description.*
Critical evaluation *Students are able to outline and communicate pros and cons of their chosen method. Moreover, they can explain the reasoning behind these pros and cons. [20 pts]*	*Students give a good outline of pros and cons of a chosen method. They are able to effectively communicate how and why these features arise, backing them up with visuals or examples highlighting the features.*	*Students give a good outline of pros and cons of a chosen method. Moreover, they can mostly communicate how and why these pros and cons arise.*	*Students give some pros and cons of the chosen method, but they are glossing over some major features/problems with the method. Alternatively, the pros and cons are simply stated without a reasoning behind them.*	*Students have not really listed pros and cons of the chosen method. Moreover, if they did it is clear that they do not understand why and how these pros and cons arise.*
Execution of method *Students show they are able to correctly execute the chosen method. They provide figures showing summarizing the output of the method as well as its performance. [20 pts]*	*Students show through visuals that they can correctly execute the chosen method. They show both visuals that can be used to answer their main network problem as well as visuals that can be used to assess the performance of the method.*	*Students show through visuals that they can correctly execute the chosen method. They only show visuals that can be used to answer their main network problem.*	*Students have executed the chosen method, but from the visuals it becomes clear that this has not been done fully correct. The visuals they show can be used to answer their main network problem, even though the conclusion will probably be flawed.*	*Students cannot show they have executed their chosen method, and even if they did it is clear that there are major flaws in the approach they took.*
Reflection on results *Students can use the results they obtained to formulate an answer to their network problem and value it. [20 pts]*	*Students use their results to answer their network problem. They show they can correctly translate from mathematical results to the real-world problem. They can also use their results to assess the validation of their approach and use it to*	*Students use their results to answer their network problem. They show they can correctly translate from mathematical results to the real-world problem. Some attempt at validating the results is done, but it is unclear how alternative*	*Students use their results to answer their network problem, even though the translation from mathematics to the real-world is a bit wonky. They do not really value their solution.*	*Students do not really use their results to answer the main network problem, and if they do then major mistakes are made in translating between mathematics and the real-world*

	formulate alternative (better) approaches.	*approaches might be derived from it.*		*setting. The solution is not valued.*
Ownership *Students show creativity through their deliverables. They do not merely execute the assignment "as is". [5 pts]*	*Students went beyond what was expected of them when executing the assignment. The poster and presentation burst with the student's own distinct "flavor".*	*Students mostly did what was expected of them, although in their analysis or executions one can sometimes see the students going beyond this.*	*Students did what was expected of them, and nothing more.*	*Students failed to do what was expected of them.*
Scientific discussion *(present.)* *Students are able to partake in scientific discussions with their peers based on their work. [5 pts]*	*During the presentations the students asked a deep question during the discussion block assigned to them. They were able to spart a good scientific discussion.*	*During the presentation the students asked a deep question during the discussion block assigned to them.*	*During the presentations the students asked a question during the block assigned to them, but this question remained on a surface-level.*	*During the presentations the students could not really ask a question in the block assigned to them.*
Verbal communication *(present.)* *Students can communicate their methods and results effectively in an oral presentation using visual aids. [5 pts]*	*The students are able to give clear presentations on their work. They effectively use visual aids, and their story is easy to follow.*	*The students are able to give a presentation on their work. They use visual aids, however at times the slides and/or story is not clear.*	*The students are able to give a presentation on their work. However, most of the time their story is not really clear.*	*The students are not able to give a clear presentation on their work. Alternatively, many parts of their story are missing.*
Written communication *(poster)* *Students can communicate their methods and results effectively in a poster. [5 pts]*	*The students are able to create a clear poster showcasing their results. They effectively switch between text, math, and visuals. It is clear what story they want to tell.*	*The students are able to create a poster showcasing their results. In the poster they lean too heavily on text, math or visuals. However, their story is mostly clear.*	*The students are able to create a poster showcasing their results. However, the story they are trying to tell is not very clear.*	*The students are not able to create a poster effectively showcasing their work. The story is not clear or many parts of it are missing.*