Exercises

1	2	3	4	5

Surname, First name

____

2AMS30 Network Statistics for Data Science
Practice Exam

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Dear student,

This is a collection of **practice questions** for the exam of the course 2AMS30 Network Statistics for Data Science. You can use this to test you understanding of the material and get a better idea of what to expect from the actual exam.

Important: the number of points associated to the different questions are not relevant and susceptible to changes in the future. So please do not pay too much attention to them.

Learning objectives
In the table on the next page you can find a table that lists for each questions which learning objectives it intents to test. You can use this to see if you mastered the different learning objectives of this course.

Question	Learning Objectives
1	Describe the network science cycle for problems related to network data.
	Execute the network science cycle on a given problem for network data.
2a	Give pros and cons of the methods for solving community detection.
2b	Interpret the results arising from community detection.
2c	Give two methods for solving the problem of community detection.
	Apply both methods to specific instances of a community detection problem.
2d	Give pros and cons of the methods for solving community detection.
3a	Give pros and cons of the methods for solving community detection.
3b	Give two methods for solving the problem of community detection.
3c	Apply both methods to specific instances of a community detection problem.
3d	Interpret the results arising from community detection.
3e	Give pros and cons of the methods for solving community detection.
	Interpret the results arising from community detection.
4a	Interpret the results arising from link prediction.
4b	Apply both methods to specific instances of a link prediction problem.
4c	Interpret the results arising from link prediction.
4d	Apply both methods to specific instances of a link prediction problem.
4e	Interpret the results arising from link prediction.
4f	Give two methods for solving the problem of link prediction.
	Give pros and cons of the methods for solving link prediction.
5a	Give two methods for solving the problem of link prediction.
5b	Give pros and cons of the methods for solving link prediction.
5c	Give pros and cons of the methods for solving link prediction.
5d	Apply both methods to specific instances of a link prediction problem.
5e	Interpret the results arising from link prediction.
5f	Apply both methods to specific instances of a link prediction problem.
5g	Interpret the results arising from link prediction.

Case study example

Decline of biodiversity is a major problem in todays world. One reason for this decline is the increase in urban environments (e.g. cities). Therefore a group of researchers wants to study how to design cities in a way that improves biodiversity. For this they want to use a network analysis approach based on the green spaces in cities and the corridors between them. They have identified several network statistics they will compute on these networks. These are:

1. Density of the network: $E/\binom{n}{2}$ (number of edges compared to the maximum possible number in simple unweighted graphs).
2. Average path length.
3. Local clustering coefficient $\frac{1}{n}\sum_{i=1}^{n}\frac{2\Delta_i}{d_i(d_i-1)}$.

Then the will compare cities based on these statistics and an overall score of biodiversity. The hope is that they will find a relation between some network structures and biodiversity which can then be taking into account when redesigning cities.

10p **1** Analyze this research setup using the Network Science Cycle. Make sure to describe the problem setting, comment on the methods and to what extend the results of the analysis can address the original research question.

↳

⤵

⤷

⤵

Modularity example question

Modularity provides a way to score a community assignment of graph. The standard formula we often see for the modularity score of a partition $\mathbf{C}_k = \{C_1, \dots, C_k\}$ for a undirected graph of size n is the following:

$$Q(\mathbf{C}_k) = \sum_{m=1}^{k} \frac{E_m}{E} - \left(\frac{D_m}{2E}\right)^2, \qquad (1)$$

where $E_m = \sum_{i,j \in C_m} A_{ij}$ and $D_m = \sum_{i \in C_m} d_i$.

1p **2a** Which of the following statements about the general modularity score is true?

□ It is based on a null model.

□ It requires knowledge on the number of communities k.

□ It is applicable to weighted networks

□ It always puts isolated components into different communities.

3p **2b** Explain the idea behind the modularity score (1). In your answer, specifically address the following two questions:

1. What makes a group of nodes a good community according to modularity?
2. Where does the term $\left(\frac{D_m}{2E}\right)^2$ come from?

In order to find communities in networks using modularity, an optimization algorithm is applied. A compact description of the algorithm is given below.

1. Assign each node to its own community
2. [Order] the nodes
3. For each node v in the assigned order do the following: a. For [each community] C check the change in modularity when adding v to C. b. Add v to the community C that yields the largest gain in modularity.
4. Create a new [graph] G' with nodes [V'] and edges [E']
5. Repeat steps 2-4 until no more gain in modularity is achieved.

5p **2c** The algorithm above has five parts with brackets []. Here some details have been omitted. Provide the missing details for each of them.

We are given a large network where nodes are researchers and links represent joint publications. Based on this data we want to find research groups, groups of people that work on similar topics. The idea being that these would be represented by communities. The network data has 10^5 researchers and 10^6 collaboration links. We also expect that the average size of a research group is around 100 with on average 500 collaboration.

3p **2d** Explain whether modularity maximization is a good approach to tackle this questions. Motivate your reasoning.

Spectral clustering example question

Spectral clustering refers to a family of methods for community detection in networks that are based on the spectrum of a matrix associated to the network.

1p **3a** Which of the following statements about spectral clustering are true?

□ It always puts isolated components into their own communities.

□ It is applicable to directed graphs.

□ It is based on a null model.

□ It requires knowledge about the number of communities.

A classical matrix used for spectral clustering is the Laplacian matrix L, whose entries are defined for unweighted graphs as

$$L_{ij} = \begin{cases} d_i - A_{ij} & \text{if } i = j, \\ -A_{ij} & \text{else.} \end{cases}.$$

However, this can be generalized to graphs with positive edge weights $w_{ij} > 0$.

1p **3b** Give the definition of the Laplacian matrix for graphs with positive edge weights.

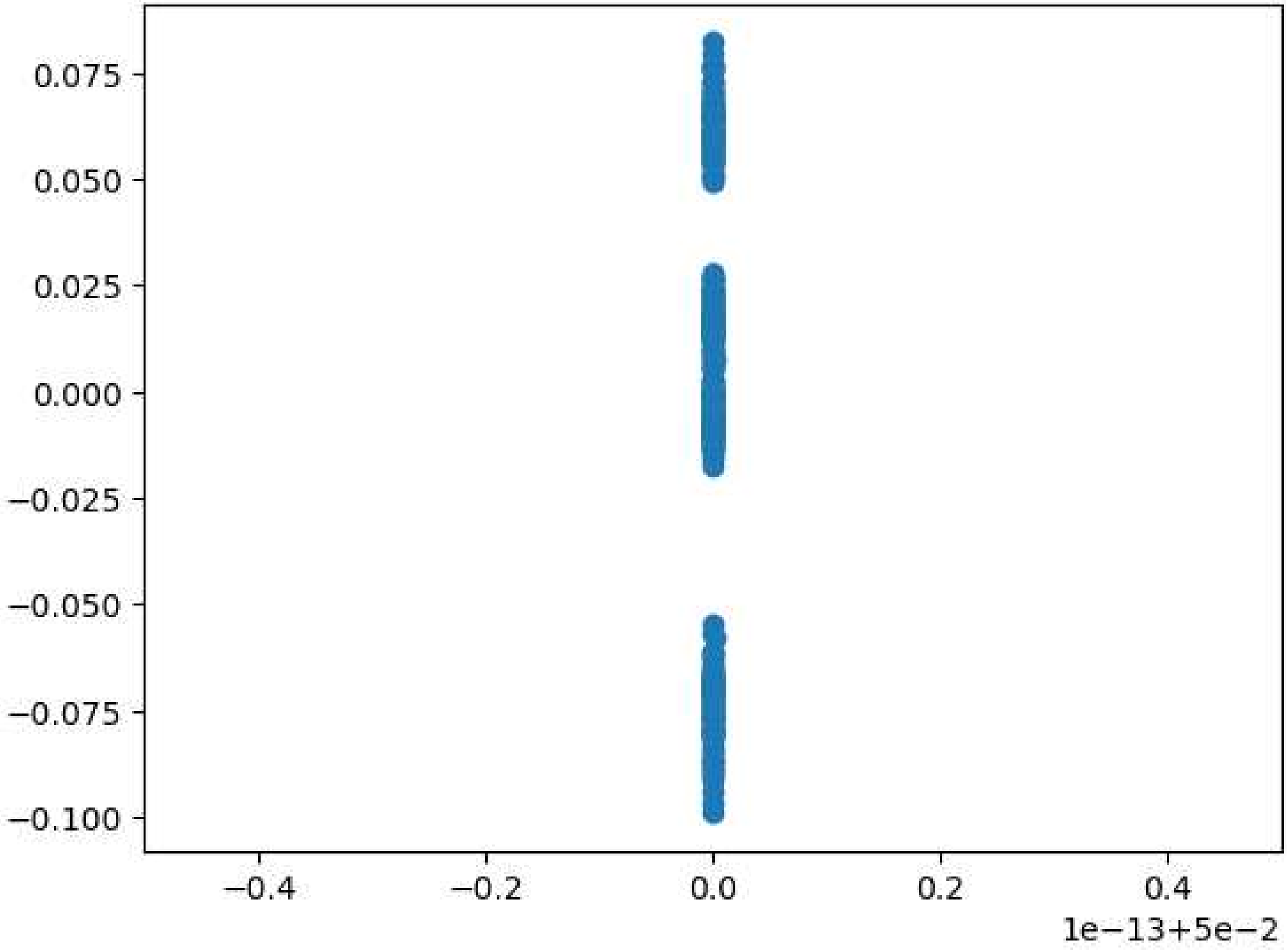
Spectral clustering with the Laplacian matrix uses the following algorithm:

1. Compute L
2. Find the smallest k eigenvalues and corresponding eigenvectors $u_1, \ldots, u_k$.
3. For each node i set $x_i = (u_{1,i}, \ldots, u_{k,i}) \in \mathbb{R}^k$.
4. Cluster the points $x_1, \ldots, x_n$ into k sets $C_1, \ldots, C_k$.

4p **3c** In the above algorithm we select the eigenvectors belonging to the *smallest* k eigenvalues. Explain in your own words why this is?

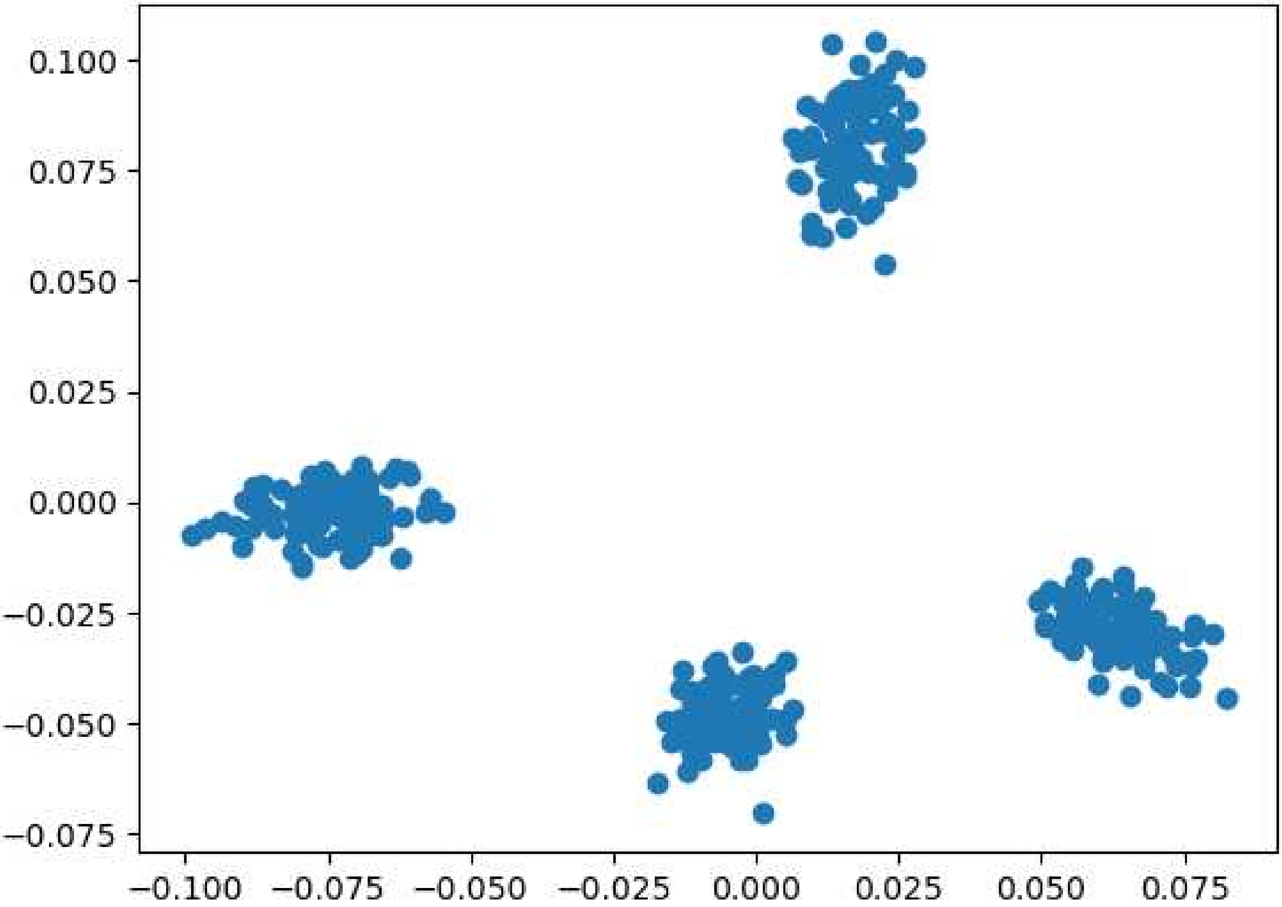
↪

Suppose we have run the above spectral clustering algorithm on a network with $k = 4$. Below you see a scatter plot of the points $(u_{1,i}, u_{2,i})$.



2p **3d** Based on this plot how many communities would you guess the network has? Explain your answer.

It turns out the true number of communities in the network is 4. We could see this from the scatter plot of the points $(u_{2,i}, u_{3,i})$ below.



3p **3e** How does the true number of communities relate to the answer of part c) and what is the problem here? Please explain your answer.

Link prediction in uniform graphs

Consider a random graph model $\mathcal{M}_n$ that creates graphs with n vertices. In this model, each (simple) graph has the same probability of being outputted. We want to do maximum likelihood based link prediction. In order to compute the reliability in this setting we will need to make an assumption about how observed graphs arise from the ground truth. Then, we can use these assumptions to compute the reliability. Suppose that an observed graphs $G^O = (V^O, E^O)$ arises from the ground truth $G = (V, E)$ with probability

$$\mathbb{P}(G^O \mid \mathcal{M}_n = G) = \alpha \cdot (|E^O| - |E|)\mathbf{1}\{G \subset G^O\}.$$

Here, α is a normalization constant independent of G and G^O. Moreover, with $G \subset G^O$ we mean that G is a strict subgraph of G^O.

2p **4a** Which of the assumptions below on the relation between G and G^O are consistent with this probability?

□ There are only missing links.

□ There are only spurious links.

□ There is no knowledge on how G^O could arise from G.

□ It is more likely that G^O differs a lot from G than a little.

□ It is more likely that G^O differs a little from G than a lot.

4p **4b** Show in this setting that the reliability for an edge $\{v, w\} \in E$ is given by

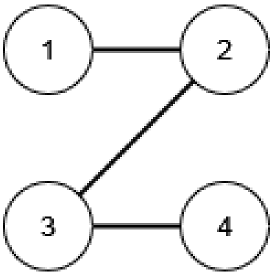
$$R(v,w) = \frac{\sum_G (|E^O| - |E|)\mathbf{1}\{G \subset G^O\}\mathbf{1}\{\{v,w\} \in E\}}{\sum_{G'} (|E^O| - |E'|)\mathbf{1}\{G' \subset G^O\}}.$$

When you look at the value of $R(v,w)$, you might notice that some of the terms in the fraction are redundant. The expression of $R(v,w)$ is easier to understand when these redundant elements are removed.

2p **4c** Do the following:

1. Highlight the terms in the expression of $R(v,w)$ that could be removed.
2. Explain why these terms can be removed.

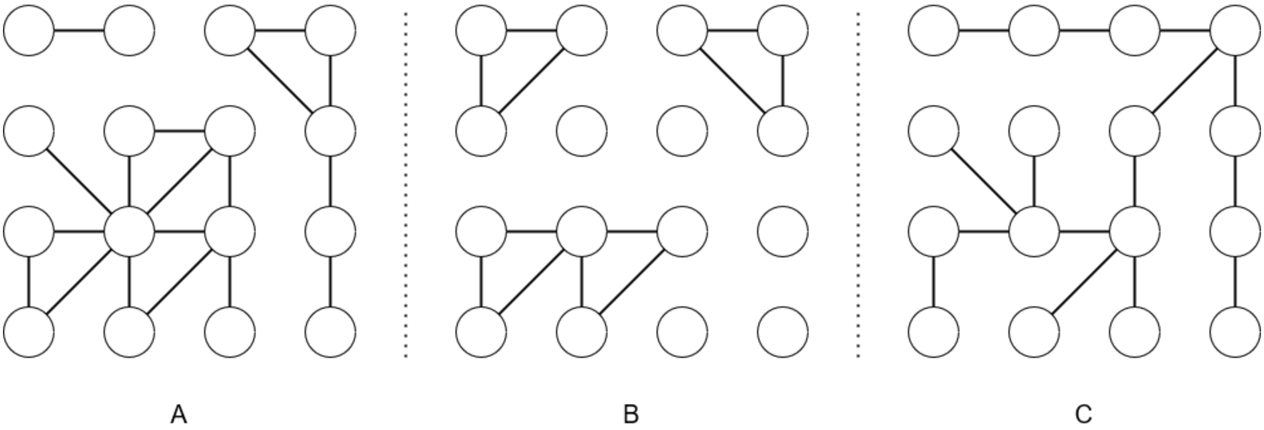
Suppose that $n = 4$ and that G^O is the graph that is given below.



3p **4d** Perform maximum likelihood link prediction with null-model $\mathcal{M}_4$ on this graph and compute for each edge its reliability.

When n becomes large, then the two sums in $R(v, w)$ become infeasible to compute. As a possible solution, you could decide to only sample "important" graphs and let these contribute to the sum.

1p **4e** Below are three graphs with $n = 16$. Assume they are all a subset of G^O. Which of these graphs will contribute the most to the edge-independent part of the reliability (i.e., is most important)?



ⓐ A ⓑ B ⓒ C

Similar to other maximum likelihood link prediction methods, you could use the Metropolis algorithm to compute $R(v, w)$.

4p **4f** Do the following:

- Explain how the Metropolis algorithm would work in this setting.
- Reflect on how this algorithm resolves the aforementioned computational infeasibility.

↳

↴

The Jaccard coefficient and link prediction

The Jaccard coefficient is a similarity index used for link prediction. If you use this index to do link prediction on an observed graph $G^O = (V^O, E^O)$, then the similarity score for two vertices $v, w \in V^O$ is given by

$$S(v,w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{|\Gamma(v) \cup \Gamma(w)|},$$

where $\Gamma(v)$ denotes the set of neighbors of vertex v.

1p **5a** Which of the following statements about the Jaccard coefficient are true?

□ It is a local similarity metric.

□ You can use it directly as a reliability score in a link prediction problem.

□ It can be used directly for link prediction on weighted graphs.

□ It is based on a null-model.

3p **5b** Explain the idea behind the Jaccard coefficient. Specifically, address the following two questions in your answer:

1. What is the intuition behind the inclusion of $|\Gamma(v) \cap \Gamma(w)|$?
2. What problem does $|\Gamma(v) \cup \Gamma(w)|$ try to solve?

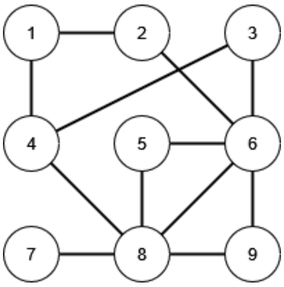
1p **5c** For which of the following types of data would the Jaccard coefficient work best as a similarity metric?

ⓐ The set of all family trees from students following 2AMS30.

ⓑ The railway network of the Netherlands.

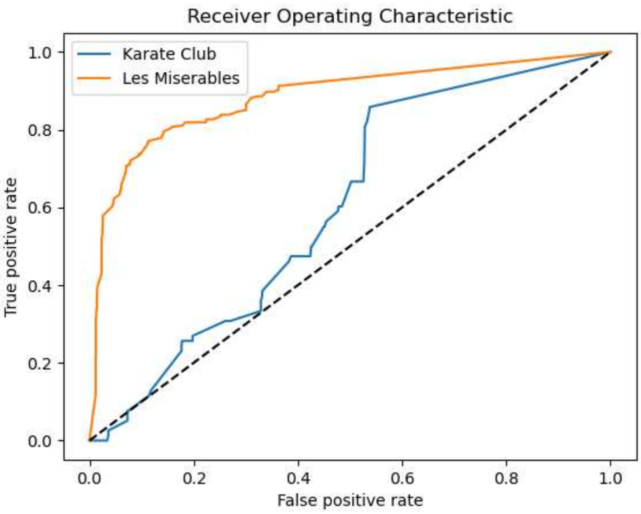
ⓒ The connections of people on Facebook

We will now use the Jaccard coefficient to do link prediction on the graph G^O below. It is given that this graph contains both missing and spurious links.



2p **5d** Compute $S(4, 6)$.

We have tested the Jaccard coefficient on two datasets to test its efficacy. For model validation we have used leave-one-out cross-validation. We have plotted the ROC-curves for both datasets below.



3p **5e** Explain the significance of the black line in the ROC-curve plot. Address the following questions in your answer:

- What does it means for an ROC-curve to be above or below this line?
- What reliability metric would always (approximately) create an ROC-curve that follows the black line?

2p **5f** Based on the ROC plot, estimate the AUC metric of the Jaccard coefficient on the "Karate Club" and "Les Miserables" dataset.

Focus on the ROC-curve of the "Les Miserables" dataset. It is given that the graph G in this dataset contains 77 nodes and 254 edges. Suppose that G^O is obtained by randomly removing one of its edges. Suppose in this setting we select the 250 links with highest Jaccard coefficient as missing.

2p **5g** Approximate the probability that this selection method will find the truly missing link.