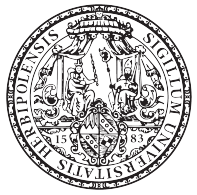

Machine Learning for Complex Networks
SoSe 2024

Prof. Dr. Ingo Scholtes
Chair of Informatics XV
University of Würzburg

Exercise Sheet 04

Published: May 29, 2024
Due: June 06, 2024
Total points: 10

Please upload your solutions to WueCampus as a scanned document (image format or pdf), a typesetted PDF document, and/or as a `jupyter` notebook.

1. **MDL with the Degreee corrected SBM**

   In Lecture 04, you learned about the Degree-Corrected Stochastic Block Model (DC-SBM). In the practice session, you saw an example of inferring communities using maximum likelihood estimation with the DC-SBM in a Jupyter Notebook. However, not specifying the number of communities can lead to trivial community assignments, as seen in Exercise 03 and Lecture 05. In Lecture 05's practice session, you learned about using the Minimum Description Length (MDL) principle to detect communities in a toy network with the standard Stochastic Block Model (SBM). Now, we will explore detecting communities in the Zachary Karate Club network using MDL with the DC-SBM.

   (a) How is the entropy S related to the number of microstates Σ of the DC-SBM? What is the probability $P(G)$ of each microstate G? [1P]

   (b) Efficient inference and description length calculation of the DC-SBM is rather involved. There are libraries that can help with this task. One such library is `graph-tool`. Refer to `https://graph-tool.skewed.de/static/doc/index.html` and use `graph-tool` to cluster the Zachary Karate Club network using the non-degree-corrected SBM and the degree-corrected SBM. What are the communities you find and their description lengths? (Tip: import `graph-tool` first before other imports to avoid conflicts with other libraries.) [3P]

   (c) Use `detect_communities_MDL` from the notebook from the practice session to cluster the Zachary Karate Club network with the standard (non-degree-corrected) SBM. What are the communities you find and their description length? Explain the difference between the description length found with `detect_communities_MDL` with the description length found using `graph-tool` with the non-degree-corrected SBM? [2P]

2. Kullback-Leibler Divergence

   For two discrete probability mass functions Q and P defined on the same sample space Ω with events i, the Kullback-Leibler divergence (which is also called relative entropy) from Q to P is defined as:

$$D_{\mathrm{KL}}(P\|Q) := -\sum_{i\in\Omega} P(i)\cdot\log\frac{Q(i)}{P(i)}$$

   (a) Consider a dice X with six faces and two different probability mass functions [2P]

$$Q(X=1)=\frac{1}{6}, Q(X=2)=\frac{1}{6}, Q(X=3)=\frac{1}{6}, Q(X=4)=\frac{1}{6}, Q(X=5)=\frac{1}{6}, Q(X=6)=\frac{1}{6}$$

and

$$P(X=1)=\frac{1}{3}, P(X=2)=\frac{1}{3}, P(X=3)=\frac{1}{12}, P(X=4)=\frac{1}{12}, P(X=5)=\frac{1}{12}, P(6)=\frac{1}{12}$$

Use `python` to compute two sequences of 1000 dice rolls with probabilities according to P and Q respectively. Use a binary Huffman code to encode the sequence and compute the number of bits required per symbol.

(b) Calculate the difference between the bits required for both sequences and compare it with the Kullback-Leibler divergence from Q to P. Interpret and explain your findings. 2P