Machine Learning for Complex Networks
SoSe 2024

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

Exercise Sheet 05

Published: June 3, 2024
Due: June 13, 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. **Random Walks and Markov chains**

   Consider a graph $G = (V, E)$ with the following adjacency matrix:

$$\mathbb{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

   (a) Assuming that the row and column numbers correspond to nodes (enumerated from 1 to 5), calculate the probability that a random walker starting in node 1 will traverse the following sequence of nodes: $(1, 2, 3, 2, 4, 1, 5, 1, 2)$. 1P

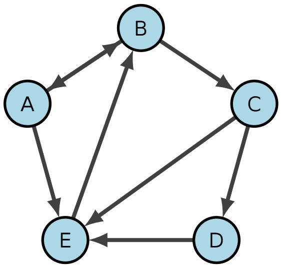
   (b) Show that for an undirected network, the stationary visitation probability of a node v converges to $\pi_v = \frac{d_v}{2 \cdot m}$ 1P

2. **Description lengths of random walks**

   (a) Generate an undirected $G(n, p)$ (no self loops, no multi-edges) and an undirected k-regular. For both networks consider a number of nodes $n = 100$. Compute the expected code length of a walk on each network. 1P
   Hint: Consider Shannon's source coding theorem.

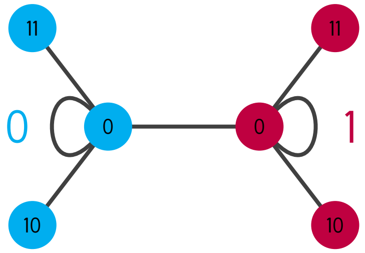
   (b) Compute the expected per-symbol code length of a random walk for a random microstate in the ensemble of undirected $G(n, p)$ random graphs (with no self loops and no multi-edges) as a function of n and p. 1P

   (c) Compute the expected per-symbol code length of a random walk on the following directed network. 1P



3. **InfoMap with exit probabilities**

Shannon's source coding theorem tells us the minimal per-symbol length of a lossless prefix-free code for the sequence of outcomes of a random variable. When the outcomes of the random variable are the nodes visited by a random walk on a network, the source coding theorem returns the minimal code length of a walk on the network. Using the MapEquation we can find an efficient code by using a hierarchical coding that utilizes the natural community structure of a network. This hierarchical code consists of two layers: In the first layer the source coding theorem is used to compute the minimal code length for the transitions between communities. In the second layer the source coding theorem is used in each community separately to compute the minimal code length for the node transitions within the community. The split of the network in communities allows us to *reuse* the shorter code-words for the nodes of different communities and for the transition between communities. However, the simplified MapEquation described in L06 does not ensure that the resulting code is unambiguous. In the following network, nodes and communities have been named using the definition of the MapEquation given in Lecture L06.



It is easy to see that a sequence of codewords does not uniquely identify a walk, i.e. is ambiguous. For example, the sequence 0010 can either refer to the walk $\boxed{0}\,010$ or to the walk $\boxed{0}\boxed{0}\boxed{1}0$.

In the full formulation of the MapEquation this issue is avoided by assigning an additional *exit codeword* to each community. The exit codeword signals that the current codeword is the last one from the current community, which implies that the next codeword will identify a new community. This procedure separates the codewords of nodes of different communities and uniquely identifies a walk based on a sequence of codewords.

(a) Given the stationary distribution, a split in communities, and the transition matrix of a network, what is the probability $q_{i\curvearrowright}$ to exit from community i? 2P
Hint: what is the probability to traverse any of the links that exit from community i ?

(b) In the full formulation of the MapEquation the entropy of a community $C_i \subset V$ is given by: 3P

$$H(\mathcal{P}^i) = -\frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in C_i} p_\beta} \log\left(\frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in C_i} p_\beta}\right) - \sum_{\alpha \in C_i} \frac{p_\alpha}{q_{i\curvearrowright} + \sum_{\beta \in C_i} p_\beta} \log\left(\frac{p_\alpha}{q_{i\curvearrowright} + \sum_{\beta \in C_i} p_\beta}\right)$$

where $q_{i\curvearrowright}$ indicates the probability to exit community C_i and p_α is the stationary probability to visit node $\alpha \in V$. This formula uses Shannon's source coding theorem to calculate the minimal expected code length for the nodes in community C_i and for the exit from C_i.

Write down the simpler definition of $H(P_i)$ as given on slide 20 of L06, i.e. based on the entropy of visitation probabilities of nodes in cluster i. Highlight the differences in the definition of $H(P^i)$ above as compared to the simpler definition used in the lecture. Explain what the first summand in the definition of $H(P^i)$ captures. Explain the differences in the second summand compared to the simpler definition of $H(P_i)$.