Mining Massive Datasets — Final Exam

1	2	3	4	5	6	7	8	9	10	11	$\sum$
/8	/10	/11	/10	/8	/4	/8	/6	/4	/4	/12	/48

Do not write anything above this line

Name:

Student ID: Signature:

- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Pages 14-16 can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Do not unstaple the sheets!**
- Wherever answer boxes are provided, please write your answers in them.
- Please write your student ID (*Matrikelnummer*) on every sheet you hand in.
- **Only use a black or a blue pen (no pencils, red or green pens!).**
- You are allowed to use your A4 sheet of handwritten notes (two sides). **No other materials (e.g. books, cell phones, calculators) are allowed!**
- Exam duration - 90 minutes.
- This exam consists of 12 pages, 7 problems. You can earn 48 points.

1 Hidden Markov Models

Problem 1 [10 points] In this question, we will discuss hidden Markov models with **continuous** observations. We will use the notation from the lecture, where $Z_t \in \{1, ..., K\}$ denotes the state at time t and $X_t \in \mathbb{R}$ denotes the observation at time t. The conditional probability of an observation at a state k is $\Pr(X_t \mid Z_t = k, \boldsymbol{\theta}) = \mathcal{N}(X_t \mid \mu_k, \sigma_k^2)$, i.e. a Gaussian distribution parametrized by mean μ_k and variance σ_k^2. $\boldsymbol{\theta}$ is the set of parameters of the HMM, which includes the initial probabilities $\boldsymbol{\pi} \in \mathbb{R}^K$, transition probability matrix $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ and the means and variances $\{\mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2\}$.

a) Write down the log-likelihood $\log \Pr(Z_1, ..., Z_T, X_1, ..., X_T \mid \boldsymbol{A}, \boldsymbol{\pi}, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2)$ as a function of $\boldsymbol{A}, \boldsymbol{\pi}, \mu_k$'s and σ_k^2's.

$$p = p(z_1) \prod_{i=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{i=1}^{T} p(X_t \mid z_t)$$

$$\log p = \log \pi + \sum_{i=1}^{T-1} \log A_{z_t z_{t+1}} + \sum_{i=1}^{T} \log N(X_t \mid \mu_t, \sigma_t^2)$$

b) Write the forwards and backwards update equations for this HMM. Explain in a single line how they are different from the updates we studied in class (discrete observations).

$$\alpha_{t+1} = N(X_{t+1} \mid \mu_i, \sigma_i^2) \cdot \sum_{i=1}^{K} A_{ik} \cdot \alpha_t(i)$$

$$\beta_t = \sum_{i=1}^{K} A_{ik} \, N(X_{t+1} \mid \mu_i, \sigma_i^2) \cdot \beta_{t+1}(i)$$

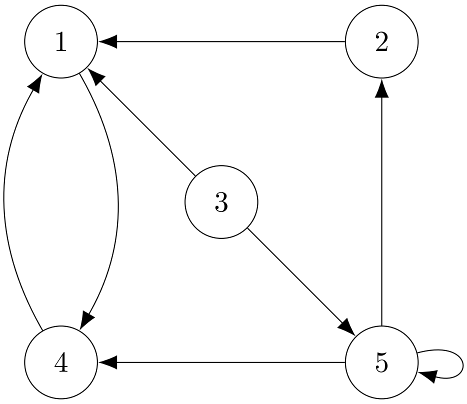
c) You are given a sequence of observations $\{X_1, ..., X_T\}$ and the corresponding states $\{Z_1, ..., Z_T\}$.

Are the maximum likelihood estimates of A_{ij} and π_i for this model different from the ones for HMM with discrete observations (that we studied in class)? Explain why or why not.

d) You are given a sequence of observations $\{X_1, ..., X_T\}$ and the corresponding states $\{Z_1, ..., Z_T\}$. Write down the closed-form maximum likelihood estimates for the parameters μ_k and σ_k^2. You don't have to derive the closed-form MLE, stating it with an intuitive explanation is enough.

2 PageRank

Problem 2 [4 points] Given the graph above and the topic-specific PageRank vector $\boldsymbol{r} = [0.3662, 0.0442, 0.0800, 0.3519, 0.1578]$ using the teleport set $S = \{3, 5\}$. What is the value of β (corresponding to the teleport probability $1 - \beta$)? Justify your answer.



3 Graph Clustering

Problem 3 [8 points] You are given a connected undirected unweighted graph with a set of nodes V. Let S_n^* be the partitioning minimizing the **normalized cut**, and let S_r^* be the partitioning minimizing the **ratio cut**. That is:

$$S_n^* = \arg\min_S \text{n-cut}(S) \qquad\qquad S_r^* = \arg\min_S \text{r-cut}(S)$$

where $S = \{C_1, V \backslash C_1\}$ is a two-way partitioning of the nodes in the graph.

Prove or disprove that $\text{n-cut}(S_n^*) \leq \text{r-cut}(S_r^*)$.

Hint: Think about the values of the cuts when we plug in the same partitioning S.

Ratio Cut: Minimize $\frac{cut(C_1,C_2)}{|C_1|} + \frac{cut(C_2,C_1)}{|C_2|}$

Normalized Cut: Minimize $\frac{cut(C_1,C_2)}{vol(C_1)} + \frac{cut(C_1,C_2)}{vol(C_2)}$

$= cut(C_1, C_2) + cut(C_1, C_1)$

$\Downarrow$

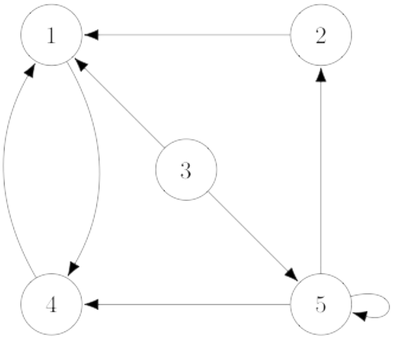
$n-cut(S) \leq r-cut(S)$

Furthermore, by definition we have that $\text{n-cut}(S_n^*) \leq \text{n-cut}(S)$. Combining these facts we obtain for any S:

$$\text{n-cut}(S_n^*) \leq \text{n-cut}(S) \leq \text{r-cut}(S)$$

Finally, replacing S with S_r^* above we obtain $\text{n-cut}(S_n^*) \leq \text{r-cut}(S_r^*)$.

Problem 2 [4 points] Given the graph above and the topic-specific PageRank vector $r = [0.3662, 0.0442, 0.0800, 0.3519, 0.1578]$ using the teleport set $S = \{3, 5\}$. What is the value of β (corresponding to the teleport probability $1 - \beta$)? Justify your answer.



$$r_5 = \beta \cdot \left(\frac{r_5}{3} + \frac{r_3}{2} \right) + (1-\beta) \cdot \frac{1}{2}$$

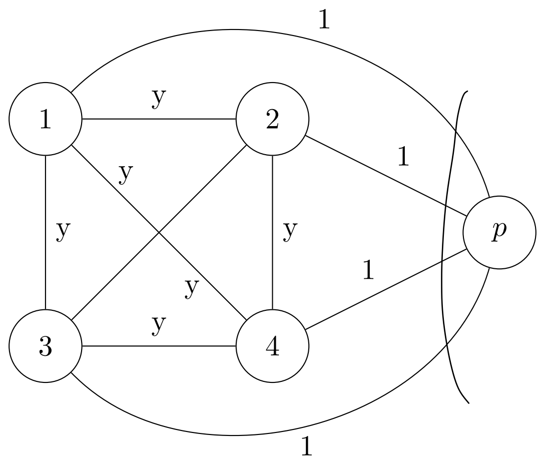
$$r_3 = (1-\beta) \cdot \frac{1}{2} = 0.08$$

$$1 - \beta = 0.16$$

$$\beta = 0.84.$$

You are given an undirected, weighted graph with a set of nodes $V = \{1, 2, \ldots, n, p\}$. It contains one clique of n nodes that is fully interconnected, each edge having weight y, and one additional node p that is connected to all nodes in the clique with weight 1.

See the example figure below for $n = 4$ as an illustration. We want to partition the nodes into **two** clusters.



a) What is the ratio cut when one of the clusters contains only the node p? Provide the solution as a function of y and n.

ratio cut $= \frac{cut(p, p\backslash V)}{|p|} + \frac{cut(p, p\backslash V)}{|p\backslash V|}$

$= \frac{n}{1} + \frac{n}{n} = n+1$

b) What is the ratio cut if one of the clusters corresponds to $S \subseteq \{1, 2, \ldots, n\}$, with $|S| = k > 0$? Provide the solution as a function of y, n, and k.

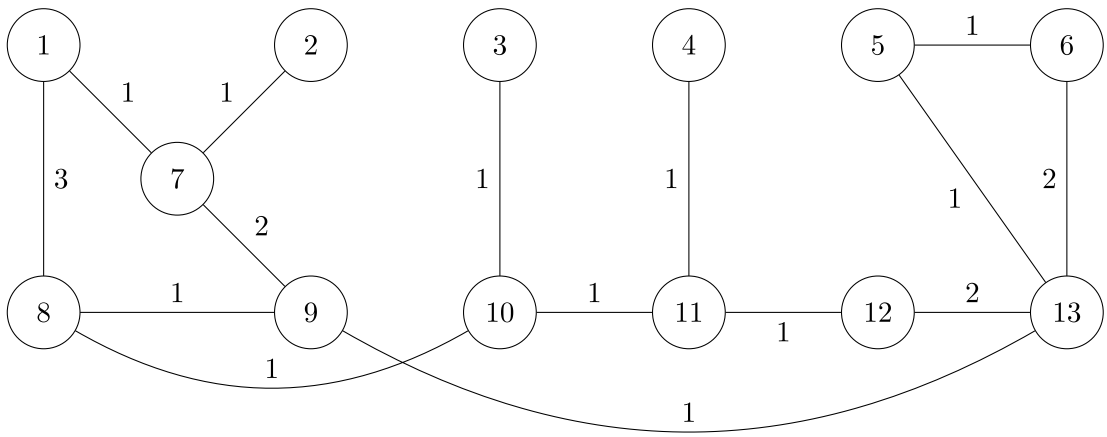
$\frac{(n-\cancel{k}\cdot k\cdot y+k}{k} + \frac{}{n-k+1}$

$y(n+1)$

c) When is the partitioning $\{p\}$ vs $\{1, 2, \ldots n\}$ the optimal solution for the minimum ratio cut? Specify all values of y and n, such that this holds.

depen on y

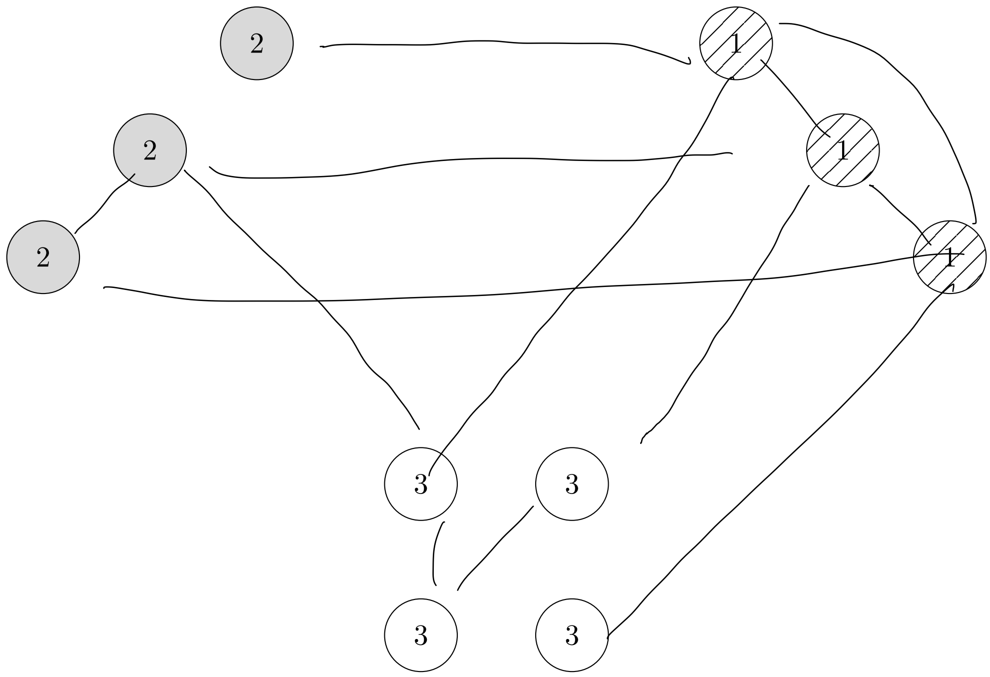
Problem 4 [6 points] Spectral embedding has been applied to the following undirected weighted graph



The plot below demonstrates the coordinates of the nodes in the embedding space (as defined by the second and the third eigenvectors of the unnormalized graph Laplacian).

Your task is to annotate the nodes in the plot below with their corresponding node IDs from the figure above.

Problem 5 [4 points] The graph below has been generated using the Stochastic block model with $K = 3$ communities (edges are hidden in the drawing). The community assignments of the 10 nodes in the graph are known (number inside each circle indicates the community ID z_i).



(a) Draw the edges in the graph such that performing maximum likelihood estimation of the parameter $\boldsymbol{\eta}$ will yield

$$\boldsymbol{\eta} = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

In case you draw some edges incorrectly and need to correct your solution, please redraw the entire graph from scratch on pages 14-16. Note that there are multiple correct solutions and specifying any of them is enough.

Provide a brief explanation below.

$$E = \begin{bmatrix} 3 & 3 & 3 \\ 3 & 1 & 4 \\ 3 & 4 & 4 \end{bmatrix}$$

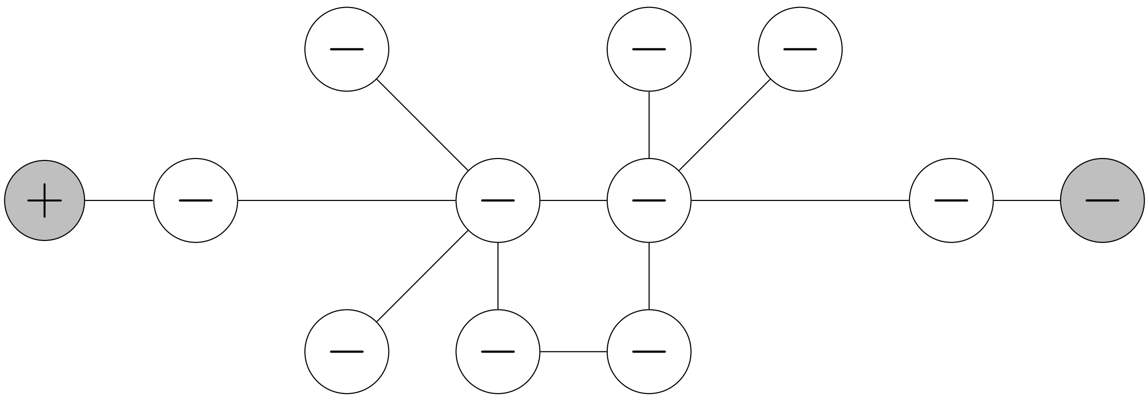
(b) What is the maximum likelihood estimate of $\boldsymbol{\pi}$ for the given graph and communities?

$\left[\frac{3}{10} \quad \frac{3}{10} \quad \frac{4}{10} \right]$

4 Label Propagation

Problem 6 [4 points] You are given the graph below with two labeled seed nodes (shaded), belonging to two classes: + and −. Specify the labels of the remaining nodes (by writing + or − inside each node) that would be obtained with an **exact** solution of the standard binary label propagation problem (i.e. assuming label smoothness).

Justify your answer. Multiple correct solutions are possible, it is enough to provide one.



winim cut

5 Deep Learning on Graphs

Problem 7 [12 points] You are given an undirected unweighted graph G with N nodes. Your task is to instantiate the differentiable message passing framework to compute PageRank. More specifically you have to specify:

a) The input features $\boldsymbol{x}_v \in \mathbb{R}^2$ of each node v.

b) The function M that computes the message from node u to its neighbor v.

c) The function U that updates the hidden representation of a node v.

such that the hidden representations in the last layer approximately recover the PageRank score of the nodes in the graph G.

Hint 1: Increasing the number of layers in this formulation should improve the approximation of the PageRank scores. Hint 2: You do not need any trainable parameters.

$$X_v = \left[\frac{1}{N}, d_v \right]$$

$$r_v = \sum \frac{r_u}{d_u}$$