Exercise 7.1: Identification of master-regulatory genes (40 points)

(a) Which dominating sets exist in the network shown above?

(b) What is the minimum dominating set (MDS) of this network?

(c) List the following sets of nodes and their sizes:
   - Largest connected component in the directed graph
   - Largest strongly connected component in the directed graph
   - Largest connected component in the underlying undirected graph

Find the minimum connected dominating set (MCDS) for each of the three sets.

(d) Compare the MDS and MCDS in terms of size and write a short conclusion.
Exercise 7.2: Co-expression based on Correlation and Mutual Information (60 points)

Mutual information ($I$) and Pearson correlation coefficient ($Corr$) between two random variables are defined as:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \times \log \left( \frac{p(x,y)}{p(x) \times p(y)} \right)$$

$$Corr(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2} \times \sqrt{\sum_{i=1}^{n} (y_i - \mu_Y)^2}}$$

where $p(x,y)$ is the joint probability distribution of expression levels $x$ and $y$, $p(x)$ is the marginal probability of expression value $x$, and $\mu_X$ is the (arithmetic) mean expression for gene $X$.

(a) Calculate the Pearson correlation coefficient and mutual information for the data given below. Here, the data is comprised of two genes whose expression were measured over 6 time series. An expressed gene is denoted by value 1. Solve this task by hand.

<table>
<thead>
<tr>
<th>Gene</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Explain the main difference between mutual information and Pearson correlation.

(c) What is the advantage of using rank-based correlation coefficients?

(d) Write a program that reads the time-series gene expression data given in the supplement and calculates the Pearson correlation coefficients for all pairs of genes.

(e) Plot the distribution of correlation coefficients between pairs of distinct genes (e.g. by using the distplot function from the Python package `seaborn`).

Interpret the shape of the plot and include it in your submission.

(f) Take a look at the correlation scores between the gene $MCTS1$ and the other genes. Write a function that finds the gene with the:

- Highest correlation to $MCTS1$
- Lowest correlation to $MCTS1$
- Correlation to $MCTS1$ that is closest to zero

Next, for each of these three genes, create a scatter plot with a linear regression model fit between its expression values and those of $MCTS1$ (e.g. with the regplot function from the Python package `seaborn`).

Include the three plots in your submission and describe what you see.