Assignment 1, Vingron Part, Complex Systems

- Paul Vogler, Mtr. Nr.:4979420

Problem 2:

Consider two random variables X and Y from which we drew the following samples:

- x = (0.3, 0.98, 0.54, 0.49, 0.39, 0.14, 0.03, 0.81, 0.65, 0.18)
- y = (0.74, 0.09, 0.48, 0.15, 0.71, 0.8, 0.53, 0.95, 0.63, 0.88)

Therefore the first observation is (x = 0.3, y = 0.74) and so on (10 observations in total). First, bin the data by dividing the interval of [0, 1] into 4 equally wide sub-intervals. Provide the following calculations by hand.

- Bin-Edges: (0, 0.25, 0.5, 0.75, 1)
- Bins (y,x): $\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix}$

A) Calculate the joint probability distribution $p_{X,Y}(x,y)$ of the binned data and write it in the following table:

Y|X	x_1	x_2	x_3	x_4
y_1	0	0.1	0	0.1
y_2	0	0	0.1	0
y_3	0.1	0.2	0.1	0
y_4	0.2	0	0	0.1

B) Calculate the marginal distributions $p_X(x)$ and $p_Y(y)$

a. $p_X(x)$: $(0.3, 0.3, 0.2, 0.2)$

b. $p_Y(y)$: $(0.2, 0.1, 0.4, 0.3)$

C) Calculate the product of the two marginal distributions pY (y) T × pX(x) (matrix multiplication!) and compare it with the joint distribution pX,Y (x, y). Are the variables X and Y stochastically independent? Justify your answer.

a. $\begin{pmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.3 \end{pmatrix} * \begin{pmatrix} 0.3 & 0.3 & 0.2 & 0.2 \end{pmatrix} = \begin{pmatrix} 0.06 & 0.06 & 0.04 & 0.04 \\ 0.03 & 0.03 & 0.02 & 0.02 \\ 0.12 & 0.12 & 0.08 & 0.08 \\ 0.09 & 0.09 & 0.06 & 0.06 \end{pmatrix}$

b. For two independent variables, the joint density is the product of their marginals. This is not the case here. Therefore, the two variables should not be independent.

D) Calculate the conditional distributions $p_{X|Y}(x|y=y_3)$ and $p_{Y|X}(y|x=x_4)$

a. $p_{X|Y}(x|y=y_3) = \frac{p_{X,Y}(x,\ y=y_3)}{p_Y(y=y_3)} = \frac{(0.1\ \ 0.2\ \ 0.1\ \ 0)}{0.4} = (0.25\ \ 0.5\ \ 0.25\ \ 0)$

b. $p_{Y|X}(y|x=x_4) = \frac{p_{X,Y}(x=x_4,\ y)}{p_X(x=x_4)} = \frac{(0.1\ \ 0\ \ 0\ \ 0.1)}{0.2} = (0.5\ \ 0\ \ 0\ \ 0.5)$

E) Calculate the joint entropy H(X, Y) and the marginal entropies H(X) and H(Y)

a. $H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) * \log_2 p(x,y) = -(6*(0.1*\log_2 0.1) + 2*(0.2*\log_2 0.2)) \approx -(6*-0.332+2*-0.464) = -(-1.992+(-0.928)) = 2.92$

b. $H(X) = -\sum p(x_i) * \log_2 p(x_i) = -\big(2*(0.3*\log_2 0.3) + 2*(0.2*\log_2 0.2)\big) \approx -(2*-0.521+2*-0.464) = -(-1.042+(-0.928)) = 1.97$

c. $H(Y) = -\sum p(y_i) * \log_2 p(y_i) = -(0.1*\log_2 0.1 + 0.2*\log_2 0.2 + 0.3*\log_2 0.3 + 0.4*\log_2 0.4) \approx -\big(-0.332+(-0.464)+(-0.521)+(-0.529)\big) = 1.846$

F) Calculate the conditional entropies H(X|Y) and H(Y|X) using the chain rule.

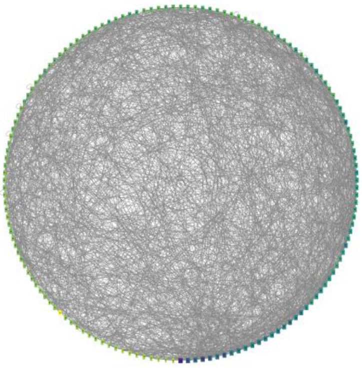
   a. $H(X|Y) = H(X,Y) - H(Y) = 2.92 - 1.846 = 1.074$
   b. $H(Y|X) = H(X,Y) - H(X) = 2.92 - 1.97 = 0.95$

G) Calculate the mutual information I(X, Y) using both, the definition and the relation to entropy. Are both results equal? Why?

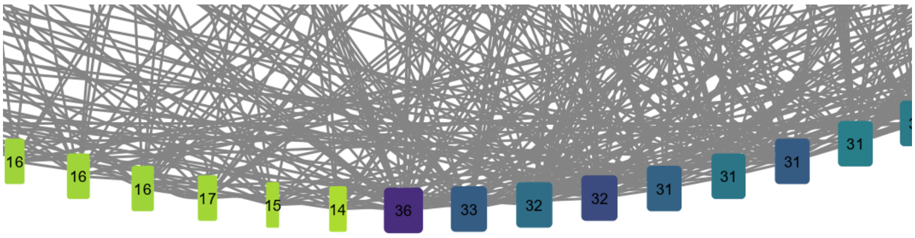
   a. $I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) * \log_2 \frac{p(x,y)}{p(x)p(y)} = 0.1 * \log_2 \frac{0.1}{0.3*0.4} + 0.2 * \log_2 \frac{0.2}{0.3*0.3} + 0.1 * \log_2 \frac{0.1}{0.3*0.2} + 0.2 * \log_2 \frac{0.2}{0.3*0.4} + 0.1 * \log_2 \frac{0.1}{0.2*0.1} + 0.1 * \log_2 \frac{0.1}{0.2*0.4} + 0.1 * \log_2 \frac{0.1}{0.2*0.2} + 0.1 * \log_2 \frac{0.1}{0.2*0.3} = 0.895$
   b. $I(X,Y) = H(X) - H(X|Y) = 1.97 - 1.074 = 0.896$
   c. $I(X,Y) = H(Y) - H(Y|X) = 1.846 - 0.95 = 0.896$
   d. Both results are basically equal, the difference just comes down to rounding precision.

Problem 3:

B) Cytoscape visualization

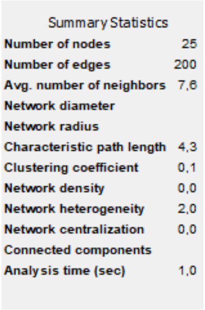


a.



b.

C) Escherichia coli interactome



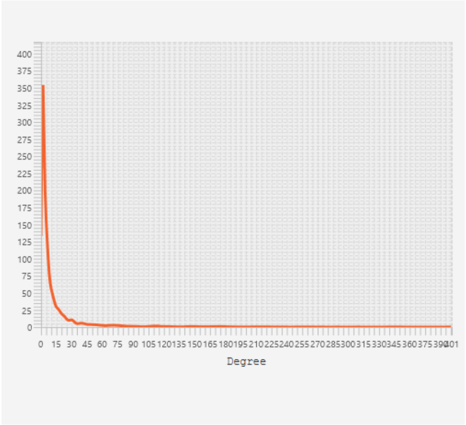
Summary Statistics

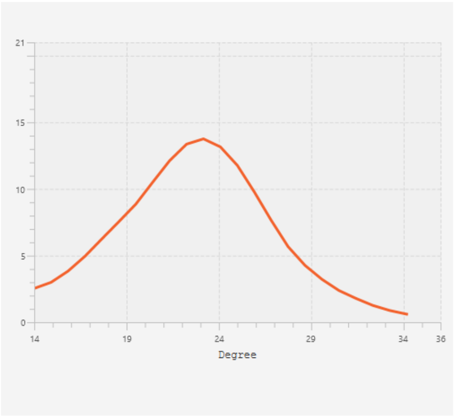
Number of nodes	25
Number of edges	200
Avg. number of neighbors	7,6
Network diameter	
Network radius	
Characteristic path length	4,3
Clustering coefficient	0,1
Network density	0,0
Network heterogeneity	2,0
Network centralization	0,0
Connected components	
Analysis time (sec)	1,0

a.

b. After the filtering of nodes without the Taxonomy ID 83333 and Nodes that do not belong to the largest connected component, there are 25 nodes and 200 edges left.

c. The left Graph shows the Degree Distribution of the E-Coli Network, while the right one shows the Degree Distribution of the Erdős-Rényi Model:





d. These two do not match, the first one is following a exponential decay distribution, while the Erdős-Rényi Model is looking like its degrees are normally distributed.