CS529 HW5

Berat Bicer - 21503050

Q1 – Key Concepts

Statement 1: A path is a walk without passing through the same link more than once.

False. A path is a walk that does not revisit past vertices as well as past edges.

Statement 2: Graph and network refer to the same thing.

False. Although they are used interchangeably, networks are real systems whereas graphs are mathematical representations.

Statement 3: The degree of a node describes the sum of its in and out degrees.

True.

Statement 4: Google pagerank is a variant of Eigenvector centrality.

True.

Statement 5: QAP analysis can be applied on any two networks.

False. The networks need to be of the same size to be comparible.

Statement 6: Hamming distance compares only binary data.

False. Hamming distance simply checks if elements at certain positions are the same and penalizes those that are different. For this reason, in string matching over large alphabets Hamming distance can still be computed.

Statement 7: Link analysis is used in law enforcement.

True.

Statement 8: Centrality metrics and rankings of nodes yield the same results regardless of whether the edge list is modeled as directed or undirected.

False. For example, degree metric is computed differently if the graph is directed.

Statement 9: Sorensen similarity gives more weight to common elements than jaccard similarity.

True.

Statement 10: Bipartite networks are only possible on two-mode networks.

False. N-mode networks where there are n unique sets of vertices may have bipartite projections.

Statement 11: All graphs with fat tailed degree distributions are scale-free graphs.

False. Many real life networks are fat-tailed although they may not be precisely following power laws statistically.

Statement 12: Network density is calculated as n(n-1)/2 where n is the number of nodes in the network.

False. The quantity is computed as a percentage of existing edges over all possible connections as L / (n * (n-1)) if directed, 2L / (n * (n-1)) if undirected.

Statement 13: Preferential attachment model works better when the limits of the nodes are well defined.

False. Since real networks grow by addition, in preferential attachment nodes have no predefined boundaries for degrees.

Statement 14: A clique of 4 nodes contains 4 3-cliques.

True.

Statement 15: The Girvan-Newman clustering algorithm is an agglomerative algorithm.

False. It’s divisive, not agglomerative.

Statement 16: Edge betweenness is a metric defined at the network level.

False. Edge betweenness is computed for a specific edge, not for the whole graph.

Statement 17: CONCOR clustering algorithm is based on structural equivalence.

True.

Statement 18: K-means requires a preset number of clusters.

True.

Statement 19: In statistics, correlation implies causation.

False. Correlation does not imply causation.

Statement 20: 99% confidence level is typically sought for in research.

True. 95% and 99% confidence intervals are popular choices.

Q2 – Pearson Coefficient

G1)

	1	2	3	4	5	6	7
1	0						
2	1	0					
3	1	0.5	0				
4	0	0.7	1	0			
5	1	0	0	1	0		
6	0	2	8	5	0	0	
7	3	0	3	0	1	0	0

$\bar{x} = \frac{\Sigma x}{n} \simeq 1.34$

Excluding self loops

From element count

$(n = 21)$

G2)

	1	2	3	4	5	6	7
1	0						
2	4	0					
3	0.3	0	0				
4	0	9	0	0			
5	0	0.1	6	7	0		
6	0	3	0	0	0	0	
7	1	0	0	0.75	0	0	0

$\bar{y} = \frac{\Sigma y}{n} \simeq 1.48$

Excluding self loops

From element count

$(n = 21)$

$$\sqrt{\sum_i (x_i - \bar{x})^2} = \sqrt{((1-\bar{x})^2 + (1-\bar{x})^2 + (0.5-\bar{x})^2 + (0-\bar{x})^2 + (0.7-\bar{x})^2 + (1-\bar{x})^2 + (1-\bar{x})^2 + (0-\bar{x})^2 + (0-\bar{x})^2 + (1-\bar{x})^2 + (0-\bar{x})^2 + (2-\bar{x})^2 + (1-\bar{x})^2 + (5-\bar{x})^2 + (0-\bar{x})^2 + (3-\bar{x})^2 + (0-\bar{x})^2 + (3-\bar{x})^2 + (0-\bar{x})^2 + (1-\bar{x})^2 + (0-\bar{x})^2)}$$

$$= 8.93$$

$$\sqrt{\sum_i (y_i - \bar{y})^2} = \sqrt{(4-\bar{y})^2 + (0.3-\bar{y})^2 + 12(0-\bar{y})^2 + (9-\bar{y})^2 + (0.1-\bar{y})^2 + (6-\bar{y})^2 + (7-\bar{y})^2 + (3-\bar{y})^2 + (1-\bar{y})^2 + (0.75-\bar{y})^2}$$

$$= 12.10$$

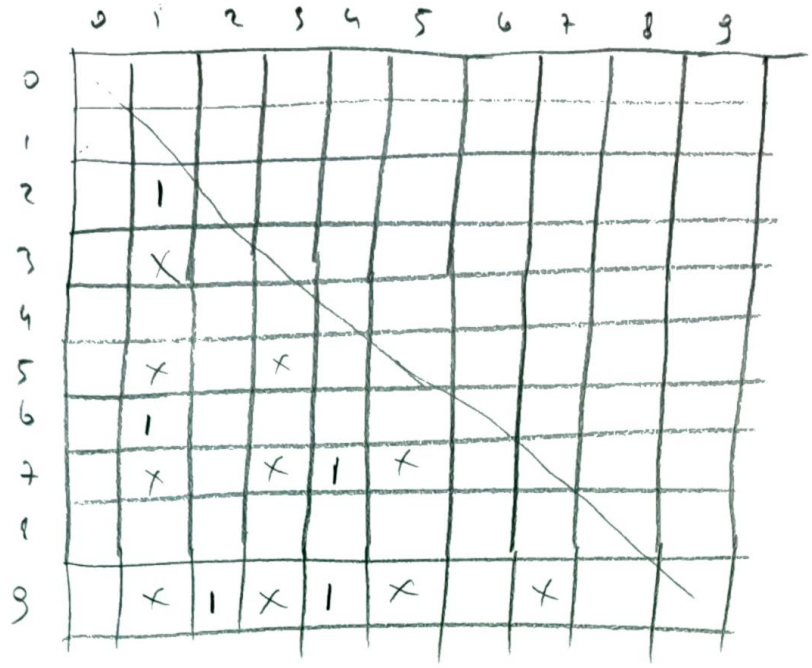
$$\sum (x-\bar{x})(y-\bar{y}) = (1-\bar{x})(4-\bar{y}) + (1-\bar{x})(0.3-\bar{y}) + (\bar{x}-0.5)(0-\bar{y}) + (0-\bar{x})(0-\bar{y})$$
$$+ (\bar{y}-1.0)(\bar{x}-0) + (\bar{y}-0)(\bar{x}-1) + (\bar{y}-0)(\bar{x}-1) + (\bar{y}-8)(\bar{x}-2.0) +$$
$$(\bar{y}-8)(\bar{x}-2) + (\bar{y}-0)(\bar{x}-0) + (\bar{y}-8)(\bar{x}-1) + (\bar{y}-9)(\bar{x}-0) +$$
$$(\bar{y}-0)(\bar{x}-0) + (\bar{y}-1)(\bar{x}-3) + (\bar{y}-0)(\bar{x}-0) + (\bar{y}-0)(\bar{x}-5) + (\bar{y}-0)(\bar{x}-8) +$$
$$(\bar{y}-0)(\bar{x}-0)(\bar{y}-0)(\bar{x}-1) + (\bar{y}-54.0)(\bar{x}-0) + (\bar{y}-0)(\bar{x}-3) +$$

$$= -15.22$$

$$\text{Pearson Coefficient} = \frac{-15.22}{8.93 \cdot 12.10} = -0.14$$

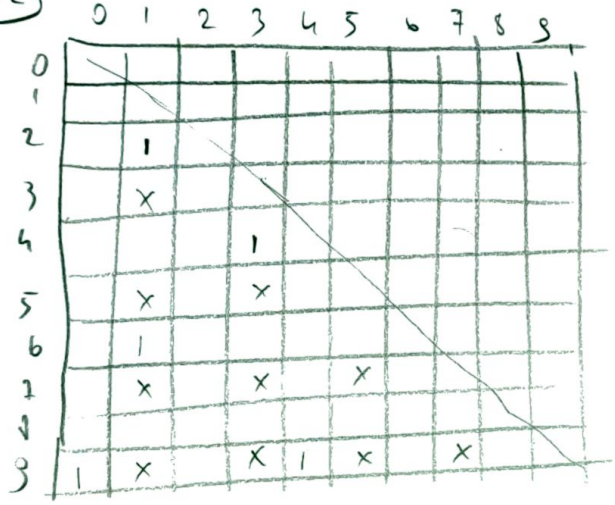
Q3 – Hamming and Jaccard Distance

61) For Ain = 0, Aout = 1, Bin = 2, Bout = 3, Cin = 4, Cout = 5, Din = 6, Dout = 7, Ein = 8 & Eout = 9,



Every other cell is zero (0), hidden for readability

62)



Every other cell is zero (0) hidden for readability

Since out-out edges are illegal these are shown with cross (x) in adjacency matrices

Ignoring the crosses,

$G_1 \rightarrow 0\,0\,1\,0\_0\,0\,0\,0\,0\,0\_0\_0\,1\,0\,0\,0\,0\,0\_0\_1\_0\,0\,0\,0\,0\,0\,0\,0\,0\,0\_1\_1\_0\_0$

$G_2 \rightarrow 0\,0\,1\,0\_0\,0\,0\,0\,1\,0\_0\_0\,1\,0\,0\,0\,0\,0\_0\_0\_0\,0\,0\,0\,0\,0\,0\,0\,0\,1\_0\_1\_0\_0$

'_' is a placeholder for illegal edges, ignored when computing similarity.

Hamming Distance = 4

$$\text{Hamming Similarity} = 1 - \frac{HD}{max\_HD} = 1 - \frac{4}{35} \simeq 0.886$$

→ max_HD: length of string ignoring placeholders since they are illegal
= 35

$$\text{Jaccard Similarity} = \frac{|G_1 \wedge G_2|}{|G_1 \cup G_2|} = \frac{31}{39} \simeq 0.794$$

$$\text{Jaccard Distance} = 1 - \text{Jaccard Similarity} \simeq 0.206$$