# Exercise Sheet 2: Social and Graph Data Management

Pablo Mollá Chárlez

## Contents

# 1 Exercise 1: Graph Measures

Consider the graph $G$ in the following figure:

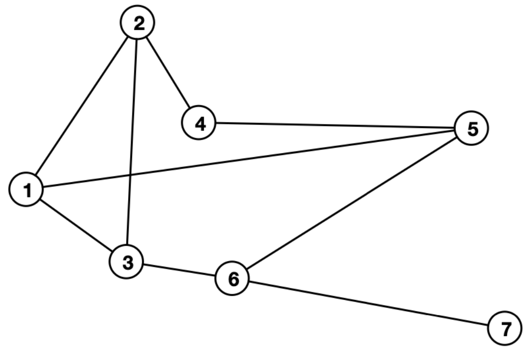


Figure 1: Graph G

- **Question 1:** Represent the graph as an adjacency list.
- **Question 2:** Write down the degree distribution of $G$, and the average degree $\langle k \rangle$.
- **Question 3:** Compute the clustering coefficient of node 1 in $G$. Explain how it is computed.
- **Question 4:** Compute the diameter $d_{max}$ of $G$, and show a path of length $d_{max}$ in $G$.
- **Question 5:** Assume that the graph was computed using a random network model with parameter $p$. What is the value of $p$? Explain how you found it

## 1.1 Answers

- **Question 1:** Represent the graph as an adjacency list.

  The adjacency list representation of the graph is:

  - $L_1 = \{2, 3, 5\}$
  - $L_2 = \{1, 3, 4\}$
  - $L_3 = \{1, 2, 6\}$
  - $L_4 = \{2, 5\}$
  - $L_5 = \{1, 4, 6\}$
  - $L_6 = \{3, 5, 7\}$
  - $L_7 = \{6\}$

- **Question 2:** Write down the degree distribution of $G$, and the average degree $\langle k \rangle$.

  The degree distribution can be extracted from the adjacency list:

  - $L_1 = \{2, 3, 5\} \implies k_1 = 3$
  - $L_2 = \{1, 3, 4\} \implies k_2 = 3$
  - $L_3 = \{1, 2, 6\} \implies k_3 = 3$
  - $L_4 = \{2, 5\} \implies k_4 = 2$
  - $L_5 = \{1, 4, 6\} \implies k_5 = 3$
  - $L_6 = \{3, 5, 7\} \implies k_6 = 3$

- $L_7 = \{6\} \implies k_7 = 1$

Therefore, the degree distribution of $G$ is:

$$p_0 = 0; \quad p_1 = \tfrac{1}{7}; \quad p_2 = \tfrac{1}{7}; \quad p_3 = \tfrac{5}{7}$$

Then, the average degree is:

$$\langle k \rangle = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 + 3 \times p_3 = \frac{18}{7}$$

- **Question 3:** Compute the clustering coefficient of node 1 in $G$. Explain how it is computed.

  The clustering coefficient of a node $i$ is computed as follows:

  $$C_i = \frac{2 \cdot e_i}{k_i \cdot (k_i - 1)}$$

  Where $e_i$ denotes the number of edges between neighbours of $i$ and $k_i$ is the degree of node $i$. Therefore, in our situation:

  $$C_1 = \frac{2 \cdot e_1}{k_1 \cdot (k_1 - 1)} = \frac{2 \cdot 1}{3 \cdot (3 - 1)} = \frac{1}{3}$$

  Here, $e_1 = 1$, because the neighoubours of node 1 are $\{2, 3, 5\}$ and only 1 edge is connecting any one of them directly (between node 2 and 3 there is an edge, but between nodes 2 and 5 there are none, as well as between 3 and 5).

- **Question 4:** Compute the diameter $d_{max}$ of $G$, and show a path of length $d_{max}$ in $G$.

  The $d_{max}$ is the longest shortest path between any two nodes in the graph, therefore the $d_{max} = 3$. One path with such length could be for instance the path that connects node 2 and node 7, which is:

  $$2 \underbrace{\longrightarrow}_{1} 3 \underbrace{\longrightarrow}_{1} 6 \underbrace{\longrightarrow}_{1} 7$$

  Another path with maximum distance could be:

  $$1 \underbrace{\longrightarrow}_{1} 3 \underbrace{\longrightarrow}_{1} 6 \underbrace{\longrightarrow}_{1} 7$$

- **Question 5:** Assume that the graph was computed using a random network model with parameter $p$. What is the value of $p$? Explain how you found it.

  The theory studied during the lectures tells us that the average degree $\langle k \rangle$ in a random network with size $N$ nodes, is described as:

  $$\langle k \rangle = p \cdot (N - 1) \longleftrightarrow p = \frac{\langle k \rangle}{N - 1} = \frac{\frac{18}{7}}{7 - 1} = \frac{18}{42} = \frac{9}{21} = \frac{3}{7}$$

# 2 Exercise 2: Uncertain Graphs and Influence

Consider the probabilistic (or uncertain) graph $\mathcal{G}$ in the following figure, where each edge is annotated with its independent probability of existing:

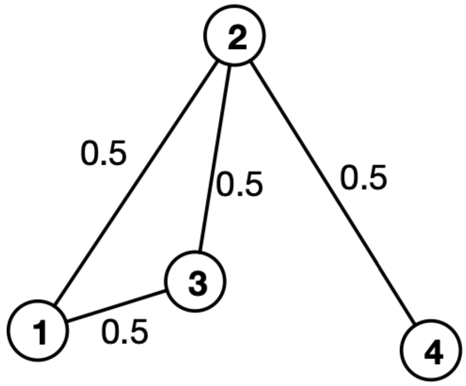


Figure 2: Graph $\mathcal{G}$

- **Question 1:** Give a possible world G of the graph $\mathcal{G}$ (i.e., a deterministic graph of 4 nodes resulting from $\mathcal{G}$). How would you compute the probability of $\mathcal{G}$?
- **Question 2:** Compute the reachability probability between nodes 1 and 4. Explain how you obtained it.
- **Question 3:** Compute the expected influence of node 1 under the influence cascade model. Explain how you obtained it.

## 2.1 Answers

Let's work step by step to address each question based on the given probabilistic graph.

- **Question 1:** Give a possible world G of the graph $\mathcal{G}$ (i.e., a deterministic graph of 4 nodes resulting from $G$). How would you compute the probability of $\mathcal{G}$?

  A possible world is a deterministic version of the graph $\mathcal{G}$ where every edge either exists or does not exist. Each edge is determined independently based on its probability. Therefore, a possible world could be the following graph:

  - **Edge $(1,2)$:** exists
  - **Edge $(2,3)$:** does not exist
  - **Edge $(1,3)$:** exists
  - **Edge $(2,4)$:** exists

  This deterministic graph $G$ includes edges $(1,2)$, $(1,3)$, and $(2,4)$, but excludes $(2,3)$. The probability of $G$ is calculated by multiplying:

  1. The probabilities of the edges that exist in $G$.
  2. The complement of the probabilities (i.e., $1-p$) for edges that do not exist.

  Then, in our case:

$$P(G) = P(1,2) \cdot (1 - P(2,3)) \cdot P(1,3) \cdot P(2,4) = 0.5 \cdot (1 - 0.5) \cdot 0.5 \cdot 0.5 = 0.5^4 = 0.0625$$

- **Question 2:** Compute the reachability probability between nodes 1 and 4. Explain how you obtained it.

  The reachability probability between nodes 1 and 4 is the probability that there exists at least one path connecting node 1 to node 4 in the probabilistic graph. There are two possible paths connecting 1 and 4 in $\mathcal{G}$:

  1. $\boxed{1 \to 2 \to 4} \implies$ Edges $(1,2),(2,4)$ must exist $\implies P(1 \to 2 \to 4) = 0.5 \cdot 0.5 = 0.25$.
  2. $\boxed{1 \to 3 \to 2 \to 4} \implies$ Edges $(1,3),(3,2),(2,4)$ must exist $\implies P(1 \to 3 \to 2 \to 4) = 0.5 \cdot 0.5 \cdot 0.5 = 0.125$.

  Combining both probabilities:

  $$P(1 \to 4) = P(1 \to 2 \to 4) + P(1 \to 3 \to 2 \to 4) = 0.25 + 0.125 = 0.375$$

  DOUBT: The two paths are disjoint, so their probabilities can be summed or as they share the common path $2 \to 4$ we should substract the overlap and how to compute the overlapping probability path?

- **Question 3:** Compute the expected influence of node 1 under the influence cascade model. Explain how you obtained it.

  The expected influence of node 1 is the expected number of nodes that can be reached from node 1 in the probabilistic graph under the influence cascade model.

  The probability of reaching each node from 1:

  - **Node 1:** Always reachable ($P = 1$).
  - **Node 2:** Reachable directly via $(1,2)$ ($P = 0.5$).
  - **Node 3:** Reachable directly via $(1,3)$ ($P = 0.5$).
  - **Node 4:** Reachable via paths $1 \to 2 \to 4$ or $1 \to 3 \to 2 \to 4$ ($P = 0.375$, as computed earlier).

  The expected influence is the sum of reachability probabilities:

  $$\text{Expected Influence} = P(1) + P(2) + P(3) + P(4) = 1 + 0.5 + 0.5 + 0.375 = 2.375$$

# 3 Exercise 3: Link Prediction

Consider again the graph $G$ in Exercise 1. Consider node 5 in the graph, and observe that the edges $(5,2)$, $(5,3)$ and $(5,7)$ are missing. We want to predict the next link from node 5 by taking, among the 3 candidate links, the one having the highest link score. We want to work only with two score functions:

1. The common neighbours score
2. The inverse distance score

Compute the link scores for each candidate link and each of the two score functions. For each of function, give the resulting best new link candidate.

## 3.1 Answers

- **Reminder:** The adjacency list for the graph:
  - $L_1 = \{2,3,5\}$
  - $L_2 = \{1,3,4\}$
  - $L_3 = \{1,2,6\}$
  - $L_4 = \{2,5\}$
  - $L_5 = \{1,4,6\}$
  - $L_6 = \{3,5,7\}$
  - $L_7 = \{6\}$
- **Common Neighbours Score:** For each candidate link $(5,x)$, the common neighbours score is computed by finding the intersection of neighbors of node 5 and node $x$.
  1. **Candidate** $(5,2)$
     - **Neighbors of 5:** $\{1,4,6\}$
     - **Neighbors of 2:** $\{1,3,4\}$
     - **Common neighbors:** $\{1,4\}$
     - **Score** $= 2$
  2. **Candidate** $(5,3)$
     - **Neighbors of 5:** $\{1,4,6\}$
     - **Neighbors of 3:** $\{1,2,6\}$
     - **Common neighbors:** $\{1,6\}$
     - **Score** $= 2$
  3. **Candidate** $(5,7)$
     - **Neighbors of 5:** $\{1,4,6\}$
     - **Neighbors of 7:** $\{6\}$
     - **Common neighbors:** $\{6\}$
     - **Score** $= 1$
- **Inverse Distance Score:** For each candidate link $(5,x)$, compute the shortest path distance $d(5,x)$ and calculate the inverse distance score as $\frac{1}{d(5,x)}$. The shortest paths from Node 5 are:
  - $d(5,2) = 2$ (Path: $5 \to 4 \to 2$) $\Longrightarrow$ Score of pair $(5,2) = \frac{1}{2} = 0.5$
  - $d(5,3) = 2$ (Path: $5 \to 6 \to 3$) $\Longrightarrow$ Score of pair $(5,3) = \frac{1}{2} = 0.5$
  - $d(5,7) = 2$ (Path: $5 \to 6 \to 7$) $\Longrightarrow$ Score of pair $(5,7) = \frac{1}{2} = 0.5$

In conclusion, if prioritizing common neighbours, we would choose $(5,2)$ or $(5,3)$ as the next link and if prioritizing inverse distance, all candidates are equally likely, and a tie-breaking criterion is needed (e.g., random selection).