60017 Tutorial: User Behaviour Graphs

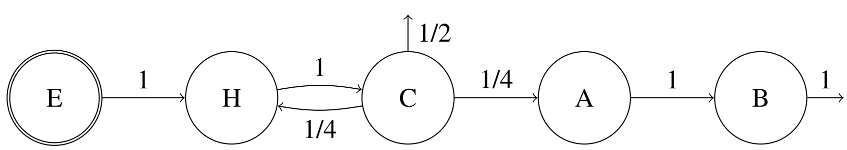
Exercise 1. A website consists of four web pages: *Home* (H), *Add* (A), *Buy* (B), and *Catalog* (C). The following user navigation sessions have been monitored in the web server log files:

- $H \to C \to A \to B$
- $H \to C \to H \to C$
- $H \to C$

Question 1.1 Draw a user behaviour graph (UBG) that models the observed sessions.
Solution: Let E and X be the dummy nodes for the entry and exit states. Then we can see the sessions as:

- $E \to H \to C \to A \to B \to X$
- $E \to H \to C \to H \to C \to X$
- $E \to H \to C \to X$

We now calculate transition probabilities, e.g., transitions out of C move into X in 50% of the cases, etc.

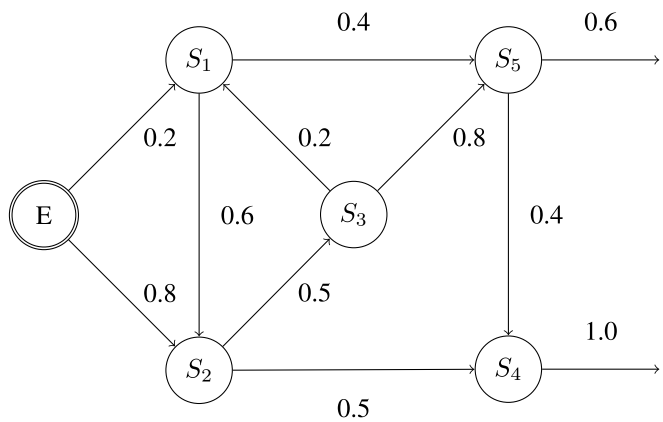


Question 1.2 Determine the visit ratio to each state of the UBG and the average session length.
Solution: From the UBG, we write the following equations:

$$
\begin{aligned}
V_E &= 1 \\
V_H &= V_E + (1/4)V_C \\
V_C &= V_H \\
V_A &= (1/4)V_C \\
V_B &= V_A
\end{aligned}
$$

Solving we find $V_H = V_C = 4/3$, $V_A = V_B = 1/3$. Therefore the average session length will be $L = V_H + V_A + V_B + V_C = 10/3 = 3.33$ pages/session.

Exercise 2. Consider the following user behaviour graph (UBG)



describing visits to pages S_1, S_2, S_3, S_4, S_5 and where E denotes the entry state.

Question 2.1 Determine the mean session length.
Solution: We begin by writing the probability transition matrix

$$P = [p_{ij}] = \begin{array}{c} \\ E \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ X \end{array} \begin{array}{c} \begin{array}{ccccccc} E & S_1 & S_2 & S_3 & S_4 & S_5 & X \end{array} \\ \begin{bmatrix} 0 & 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \end{array}$$

The visits are obtained from the system of linear equations

$$\begin{aligned} V_E &= 1 \\ V_1 &= 0.2V_3 + 0.2V_E \\ V_2 &= 0.6V_1 + 0.8V_E \\ V_3 &= 0.5V_2 \\ V_4 &= 0.5V_2 + 0.4V_5 \\ V_5 &= 0.4V_1 + 0.8V_3 \\ V_X &= 1 \end{aligned}$$

which has solutions $V_E = V_X = 1$, $V_1 = 14/47$, $V_2 = 46/47$, $V_3 = 23/47$, $V_4 = 163/235$, $V_5 = 24/47$. The mean session length is therefore $L = V_1 + V_2 + V_3 + V_4 + V_5 = \frac{698}{235} = 2.97$ pages.

Question 2.2 Give a theoretical formula to determine the probability $p_3^{(4)}$ of visiting page S_3 as fourth within the session (You are not asked to determine it numerically).
Solution: Initially $p^{(0)} = [p_E^{(0)}, p_1^{(0)}, p_2^{(0)}, p_3^{(0)}, p_4^{(0)}, p_5^{(0)}, p_X^{(0)}] = [1, 0, 0, 0, 0, 0, 0]$. Therefore we seek for the entry corresponding to $p_3^{(4)}$ in the vectors $p^{(4)} = p^{(0)}P^4$.

Question 2.3 Helping yourself with the UBG figure, can you tell from the diagram the value of $p_3^{(4)}$?
Solution: It is possible to verify numerically, for example using MATLAB, that $p_3^{(4)} = 0$. This is evident from the diagram: after 4 steps it is impossible to be in S_3. The only allowed states are only S_1, S_2 and S_5.

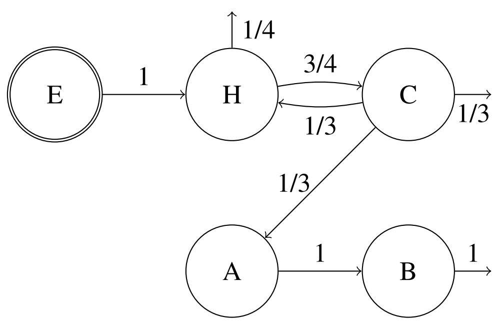
Exercise 3. A basic e-commerce website consists of the following web pages: *Home* (H), *Add* (A), *Buy* (B), and *Catalog* (C). The following user navigation sessions have been recorded in the web server log files:

- $H \to C$
- $H \to C \to A \to B$
- $H \to C \to H$

Question 3.1 Draw a user behaviour graph (UBG) that models the observed sessions.
Solution: Let E denote the entry state of the user and X that the user left, we rewrite the session as

- $E \to H \to C \to X$
- $E \to H \to C \to A \to B \to X$
- $E \to H \to C \to H \to X$

Looking at the frequency of invocation of each page, we derive the UBG as:



For example, transitions out of H are 3 times out of 4 towards C, and 1 time towards X, resulting in probabilities $3/4$ and $1/4$ out of H.

Question 3.2 Determine the visit ratio to each state of the UBG.
Solution:

$$
\begin{aligned}
V_E &= 1 \\
V_H &= (1/3)V_C + V_E \\
V_A &= (1/3)V_C \\
V_B &= V_A \\
V_C &= (3/4)V_H
\end{aligned}
$$

Substituting V_C in the expression of V_H we get $V_H = (1/4)V_H + V_E \Rightarrow V_H = \frac{4}{3}$, from which it readily follows that

$$
\begin{aligned}
V_A &= 1/3 \\
V_B &= 1/3 \\
V_C &= 1
\end{aligned}
$$

Question 3.3 Use the visit ratios to predict the average session length.
Solution:

$$L = V_H + V_A + V_B + V_C = (4/3) + (1/3) + (1/3) + 1 = 9/3 = 3$$