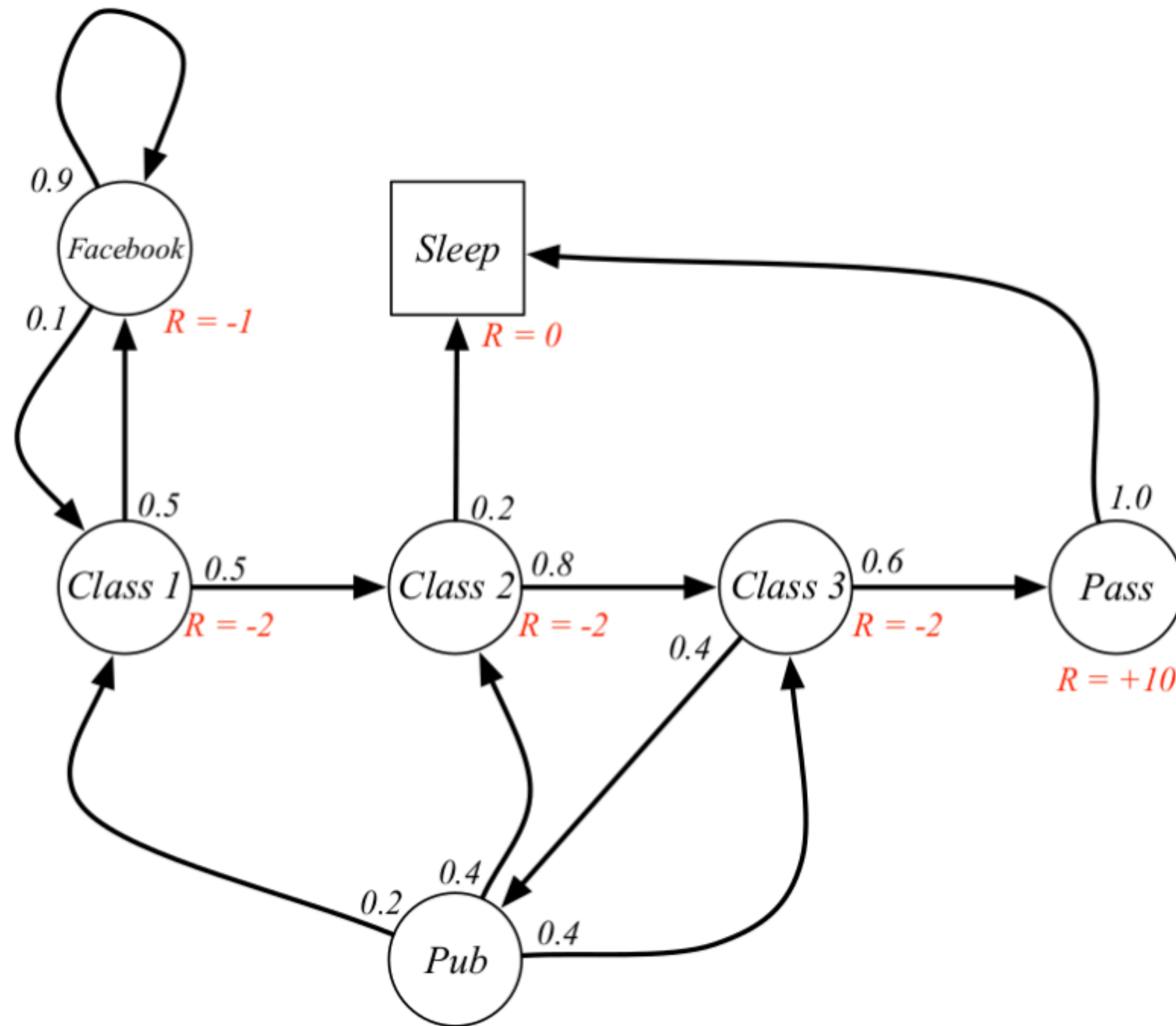


Can you explore without the past?

Readings for today

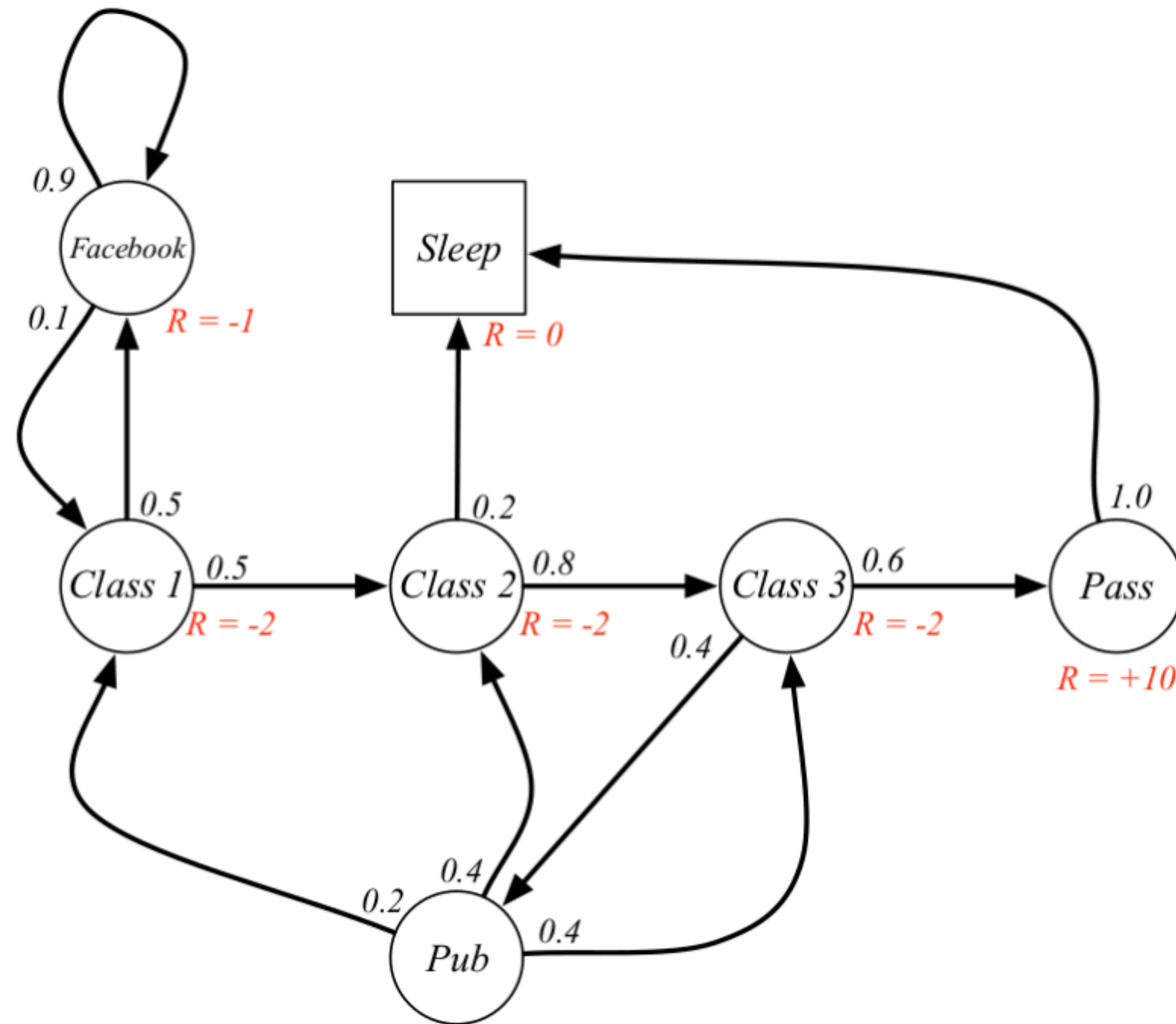
- Ashraf M. (2021). Reinforcement Learning Demystified: Markov Decision Processes (Part 1). Become Sentient.

The state-action problem



What is the best way to strategically shift from one state to another?

Markov property

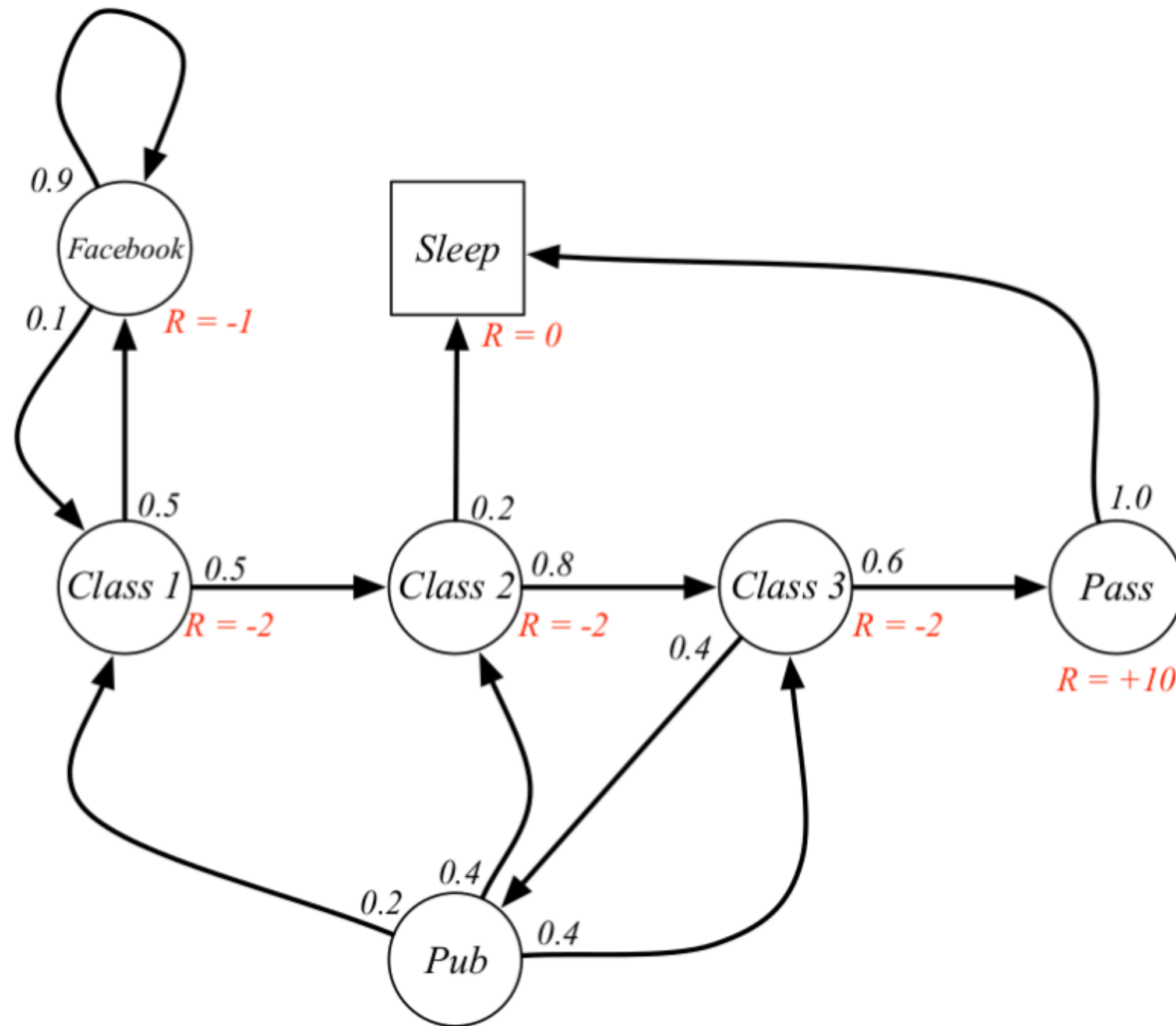


“The future is independent of the past given the present.”

State S_t has the Markov property if and only if

$$P(S_{t+1} | S_t) = P(S_{t+1} | S_1, \dots, S_t)$$

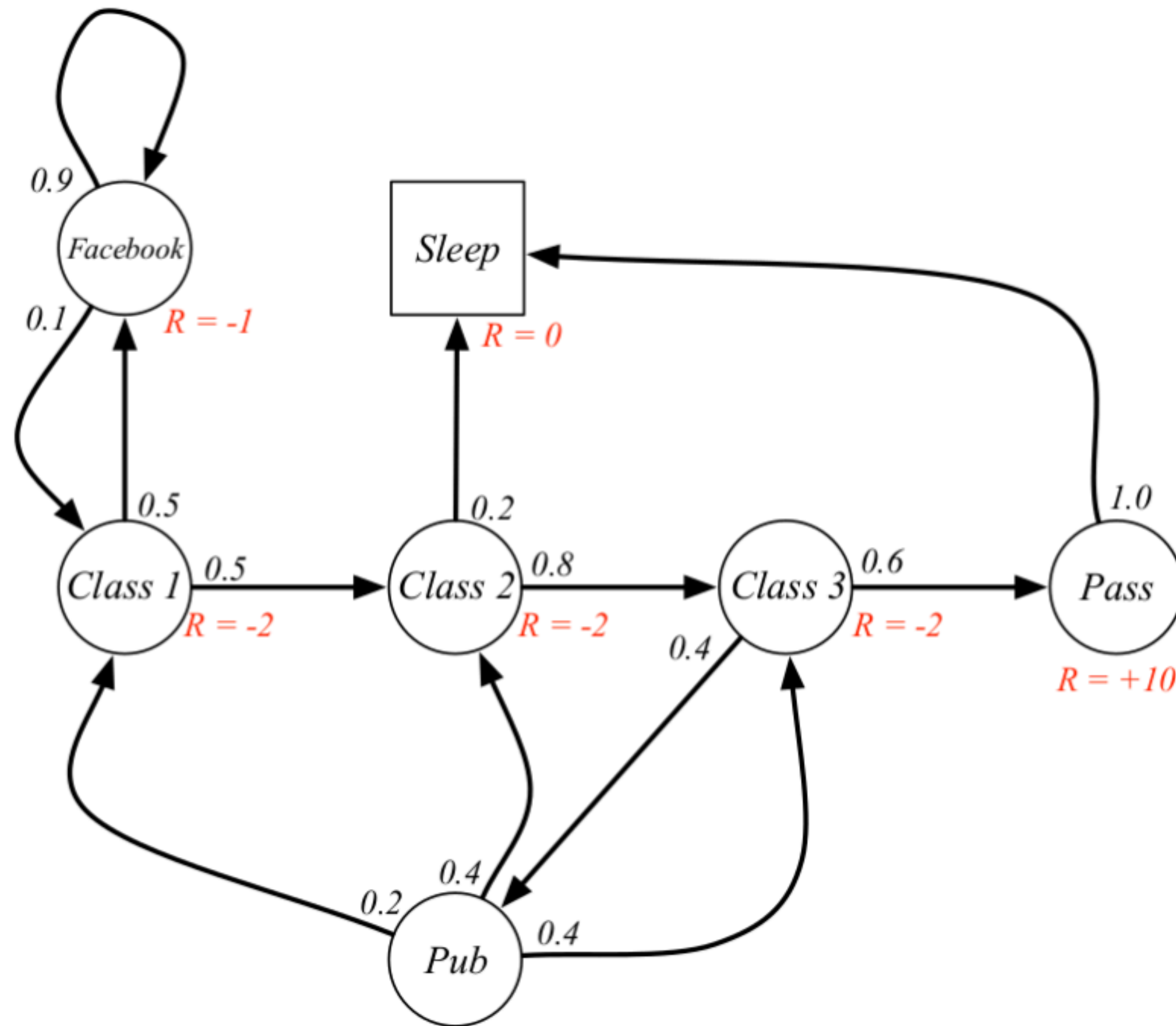
Markov property



State transition

$$\mathcal{P} = \text{from} \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \dots & \mathcal{P}_{nn} \end{bmatrix} \text{ to}$$

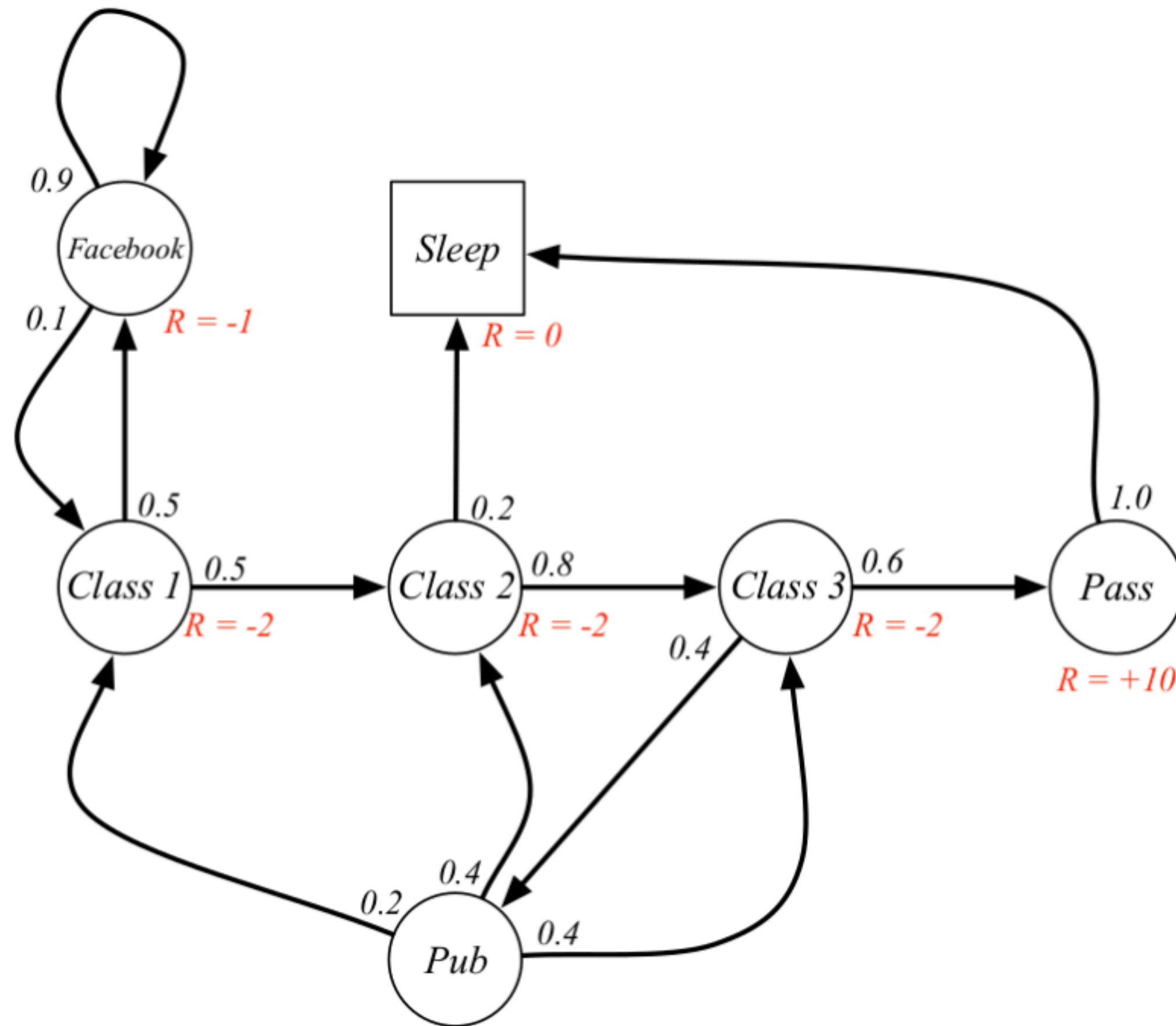
Markov process



A *Markov process* is a memory-less random process, i.e. a sequence of random states S_1, S_2, \dots with the Markov property

(\mathbf{S}, \mathbf{P})
state transition function

Markov *reward* process



A *Markov reward* process is a Markov process with a value judgement.

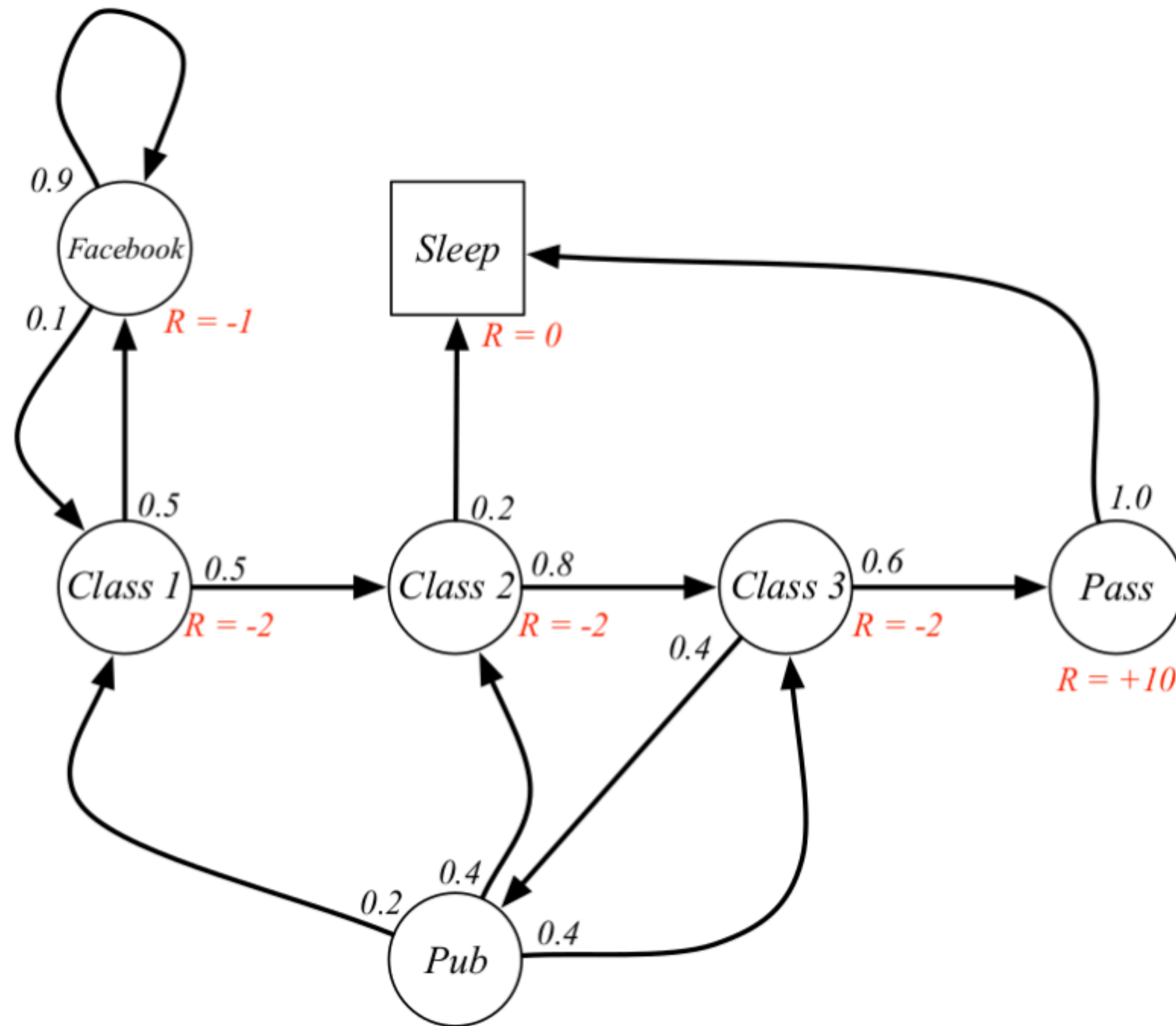
$$R_S = \mathbb{E}(R_{t+1}, S_t = S)$$

reward function discount factor

(S, P, R, γ)

state transition function

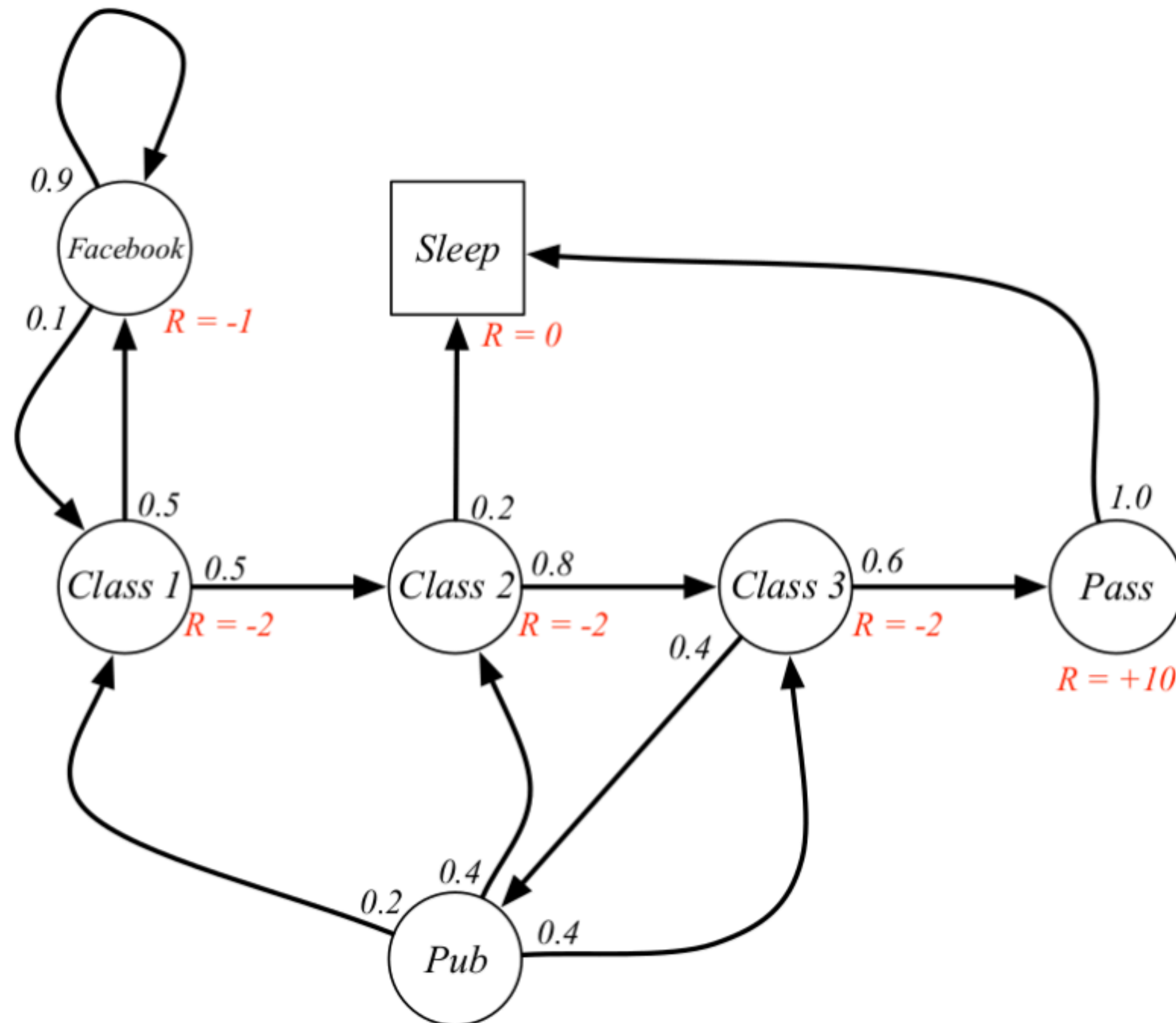
Markov *reward* process



Gain function: Far future awards are less valuable than more immediate awards.

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \end{aligned}$$

Markov *reward* process



State-value function: Expected return starting from state S

$$v(s) = \mathbb{E}(G_t | S_t = s)$$

Return for the path of [Class 1 \rightarrow Class 2 \rightarrow Class 3 \rightarrow Pass \rightarrow Sleep] is:

$$v = -2 - 2 \times \frac{1}{2} - 2 \times \frac{1}{4} + 10 \times \frac{1}{8} = -2.25$$

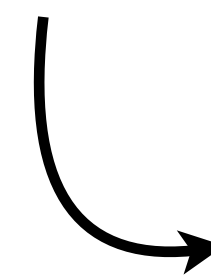
The Bellman equation

What is the *optimal* path through potential states that has the highest value?

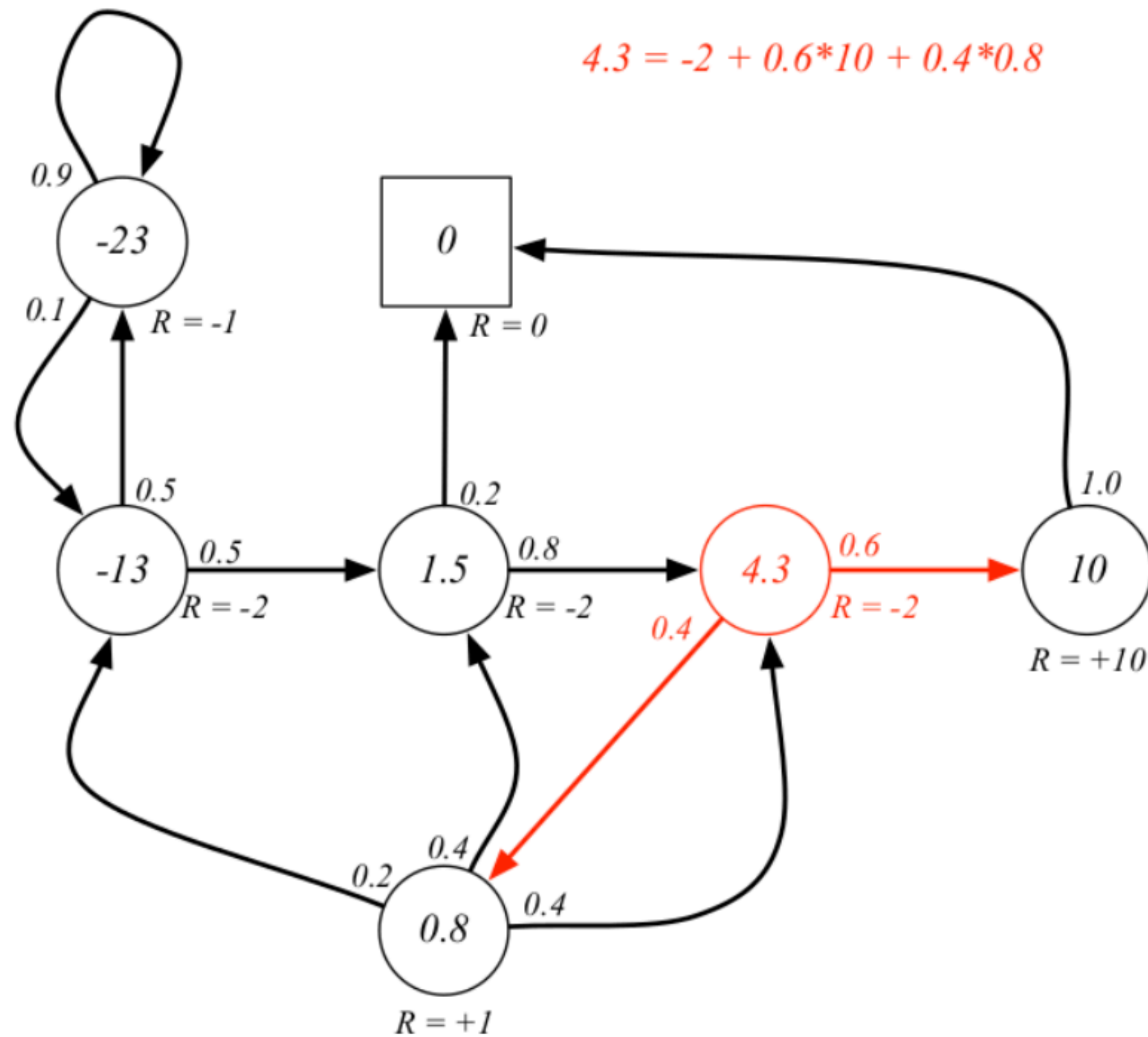
$$v(s) = \mathbb{E}(G_t | S_t = s)$$

Bellman equation

$$v(s) = \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) | S_t = s)$$

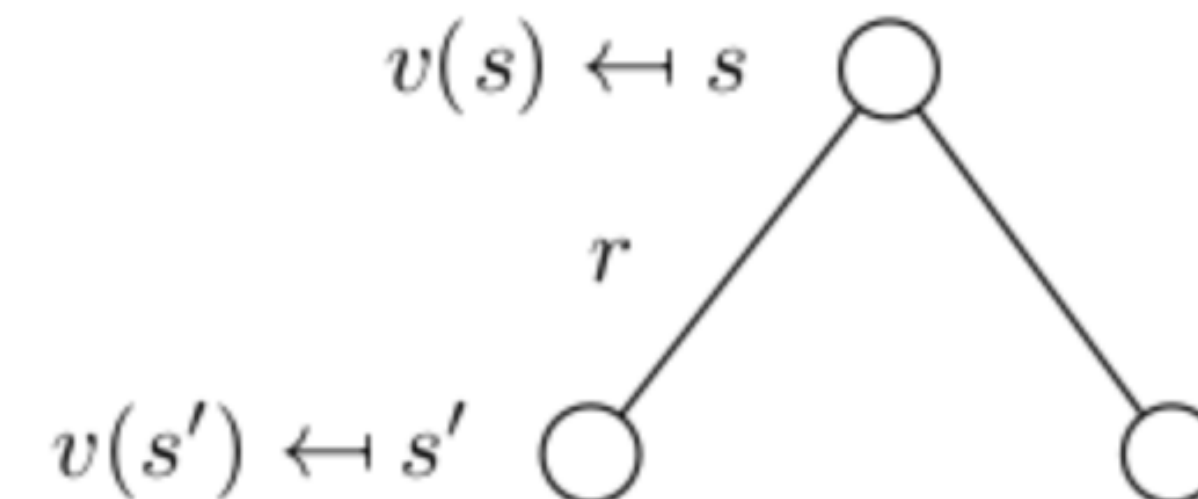

$$G_{t+1} \rightarrow v(S_{t+1})$$

The Bellman equation

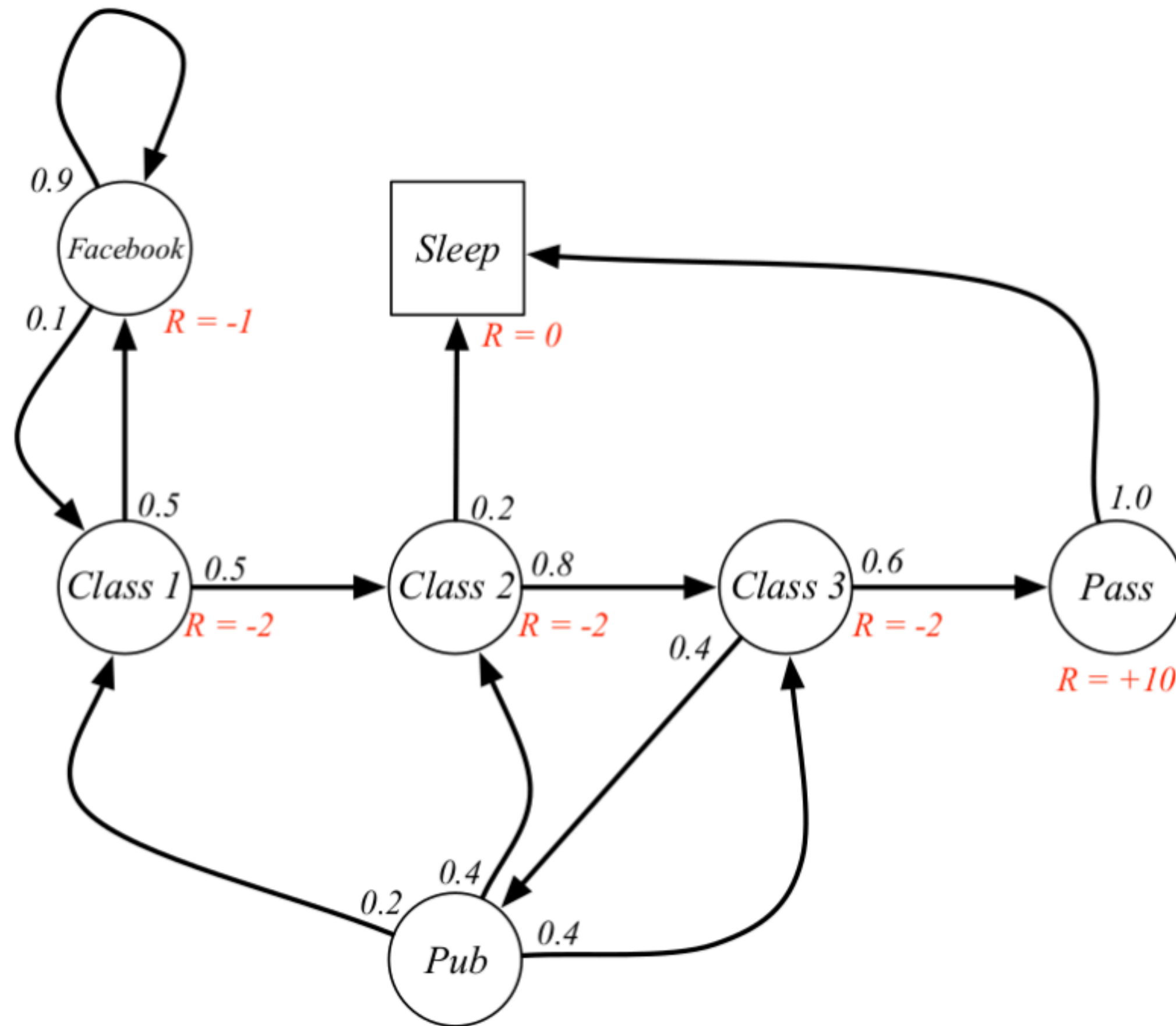


The value both depends on the reward and the transition probability .

$$\begin{aligned} v(s) &= \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s) \\ &= \mathbf{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathbf{P}_{ss'} v(s') \end{aligned}$$



Markov *decision* process

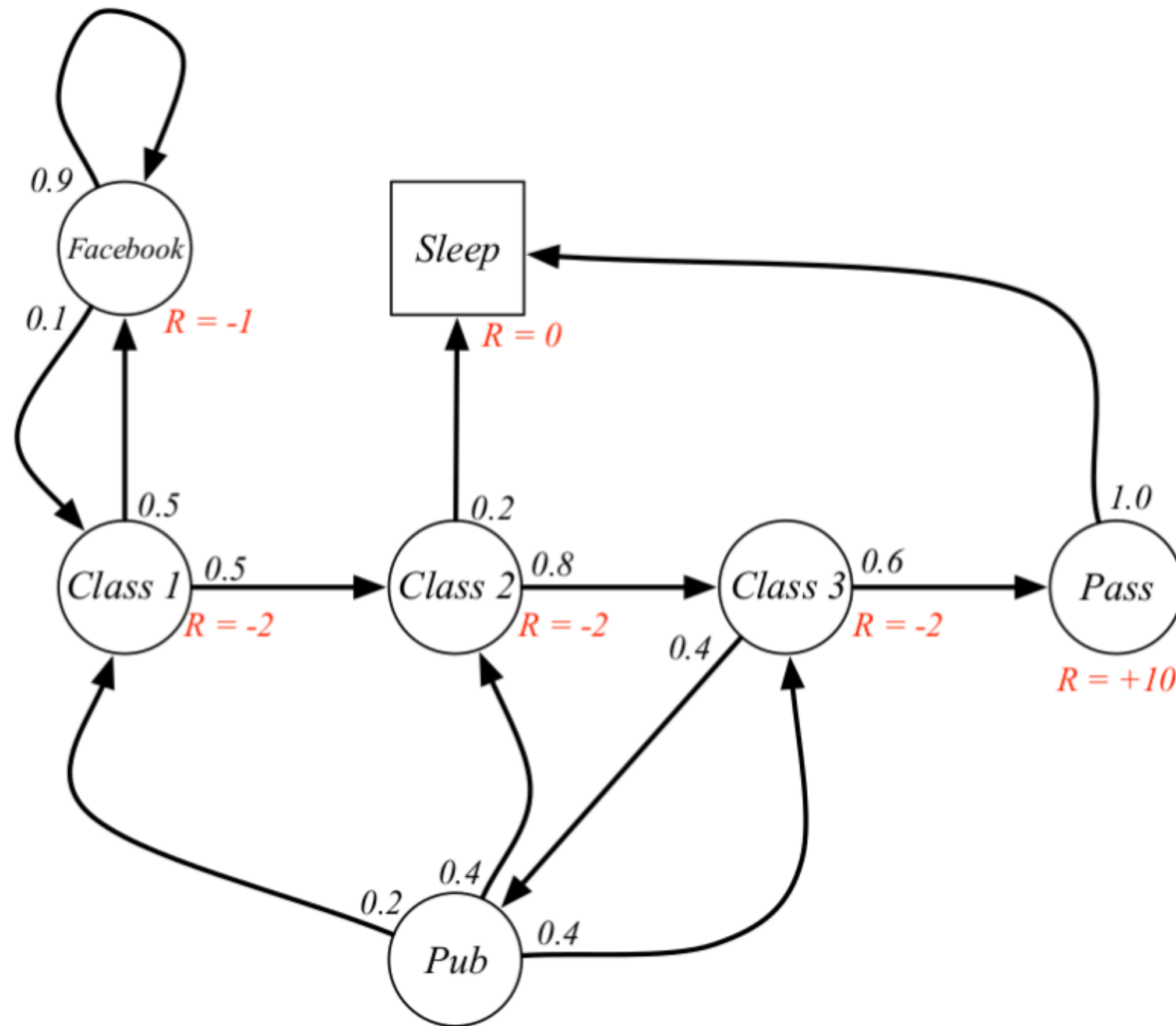


A Markov reward process, with a decision policy π .

$$\pi(a | s) = P(A_t = a | S_t = s)$$

discount factor
reward function
action
 $(\mathbf{S}, \mathbf{P}, \mathbf{R}, \gamma, \mathbf{A})$
state space
transition function

Markov *decision* process



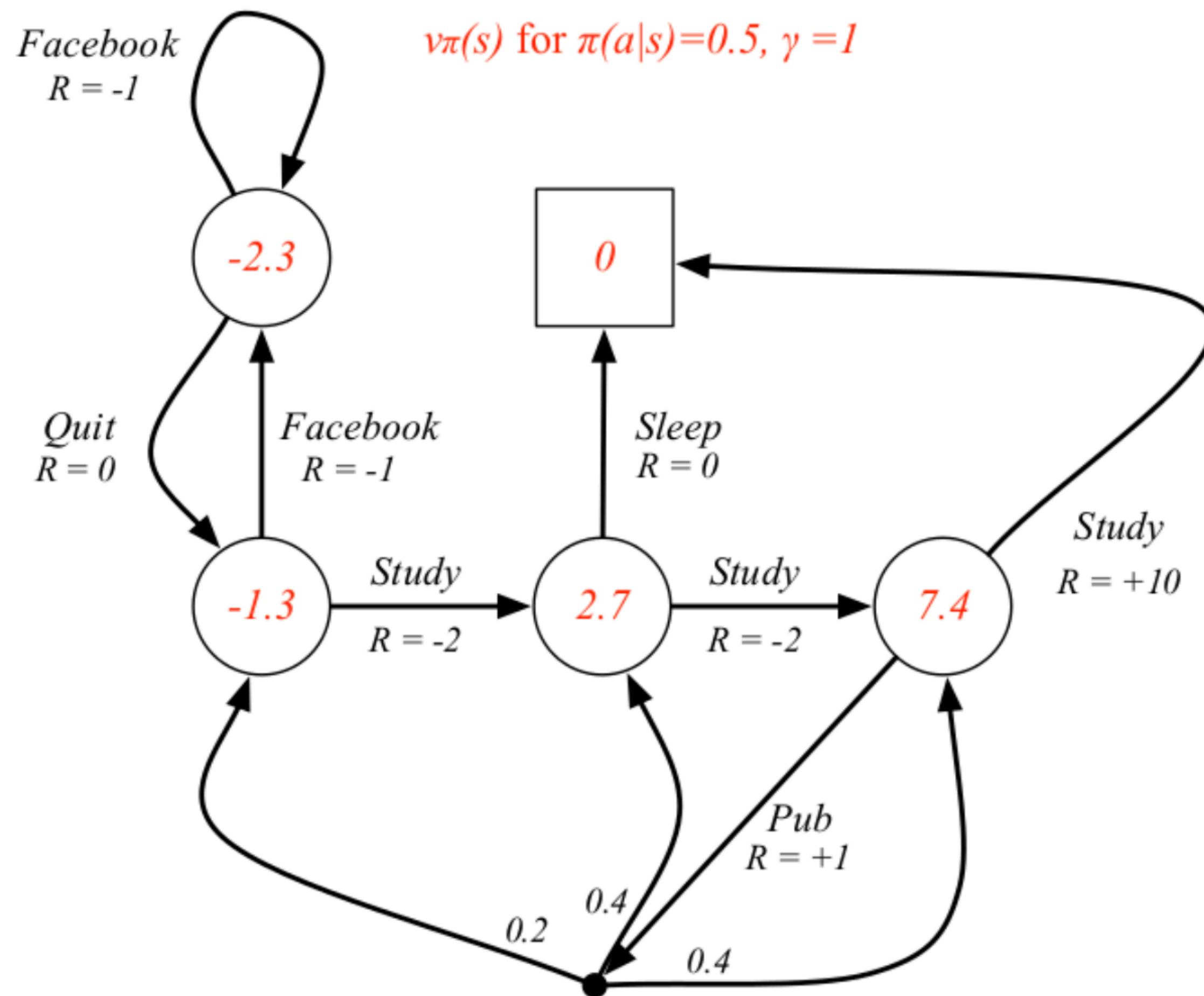
State transition function

$$P_{s,s'}^{\pi} = \sum_{a \in A} \pi(a | s) P_{s,s'}^a$$

Reward function

$$R_s^{\pi} = \sum_{a \in A} \pi(a | s) R_s^a$$

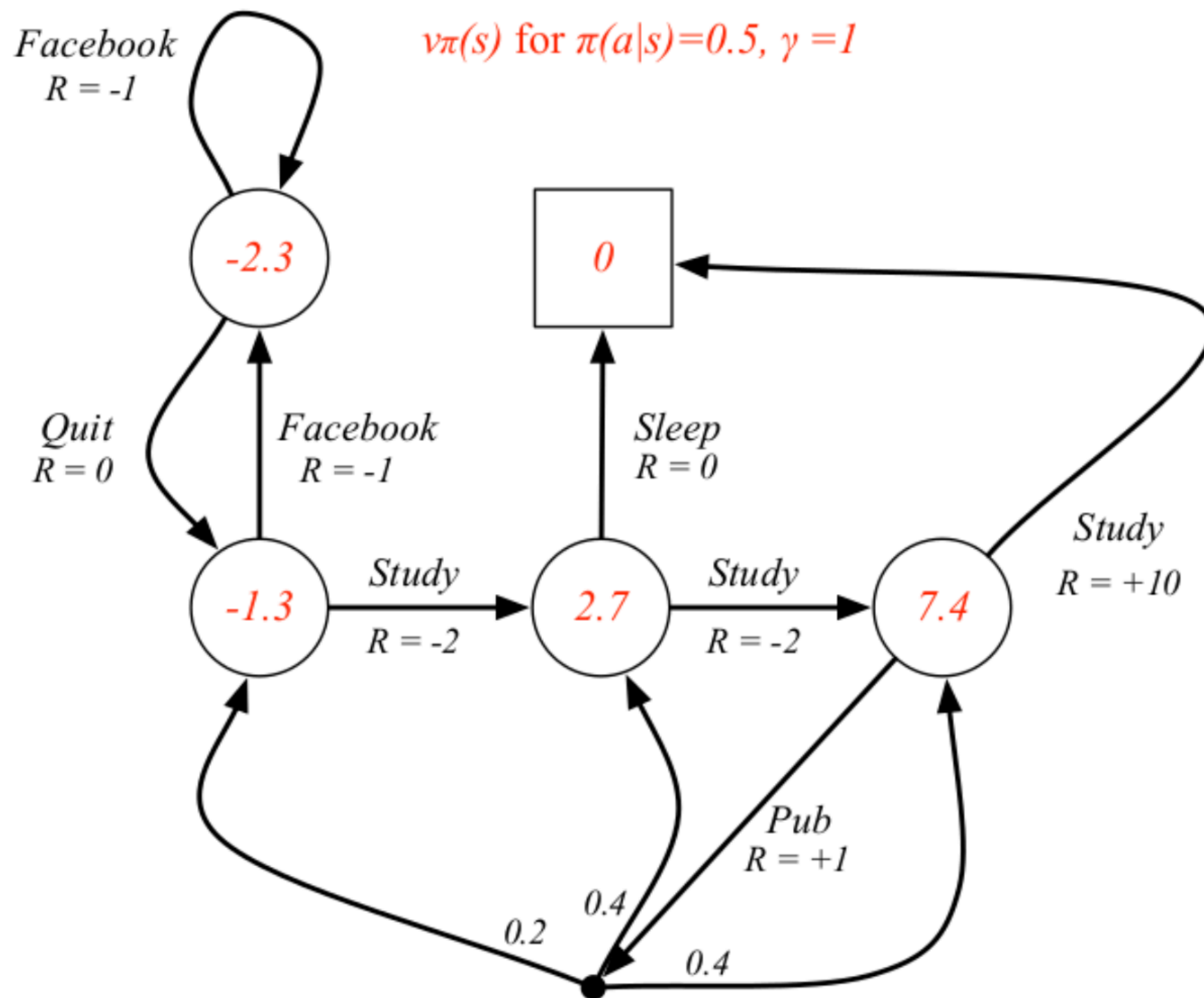
Markov *decision* process



The **state-value** function $\mathbf{V}_{\pi}(s)$ of an MDP is the expected return starting from state S , and then following policy π .

$$\begin{aligned} v_{\pi}(s) &= \mathbb{E}_{\pi}(G_t | S_t = s) \\ &= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right) \end{aligned}$$

Markov *decision* process



The **action-value** function $q_{\pi}(s, a)$ is the expected return starting from state s , taking action a , and following policy π .

$$\begin{aligned} q_{\pi}(s, a) &= \mathbb{E}_{\pi}(G_t | S_t = s, A_t = a) \\ &= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right) \end{aligned}$$

Food for thought

- Are Markov decision processes a reasonable approximation of decision-making in humans and mammals?
- Are Markov decision processes really memory-less (i.e., is the future really independent of the past)?