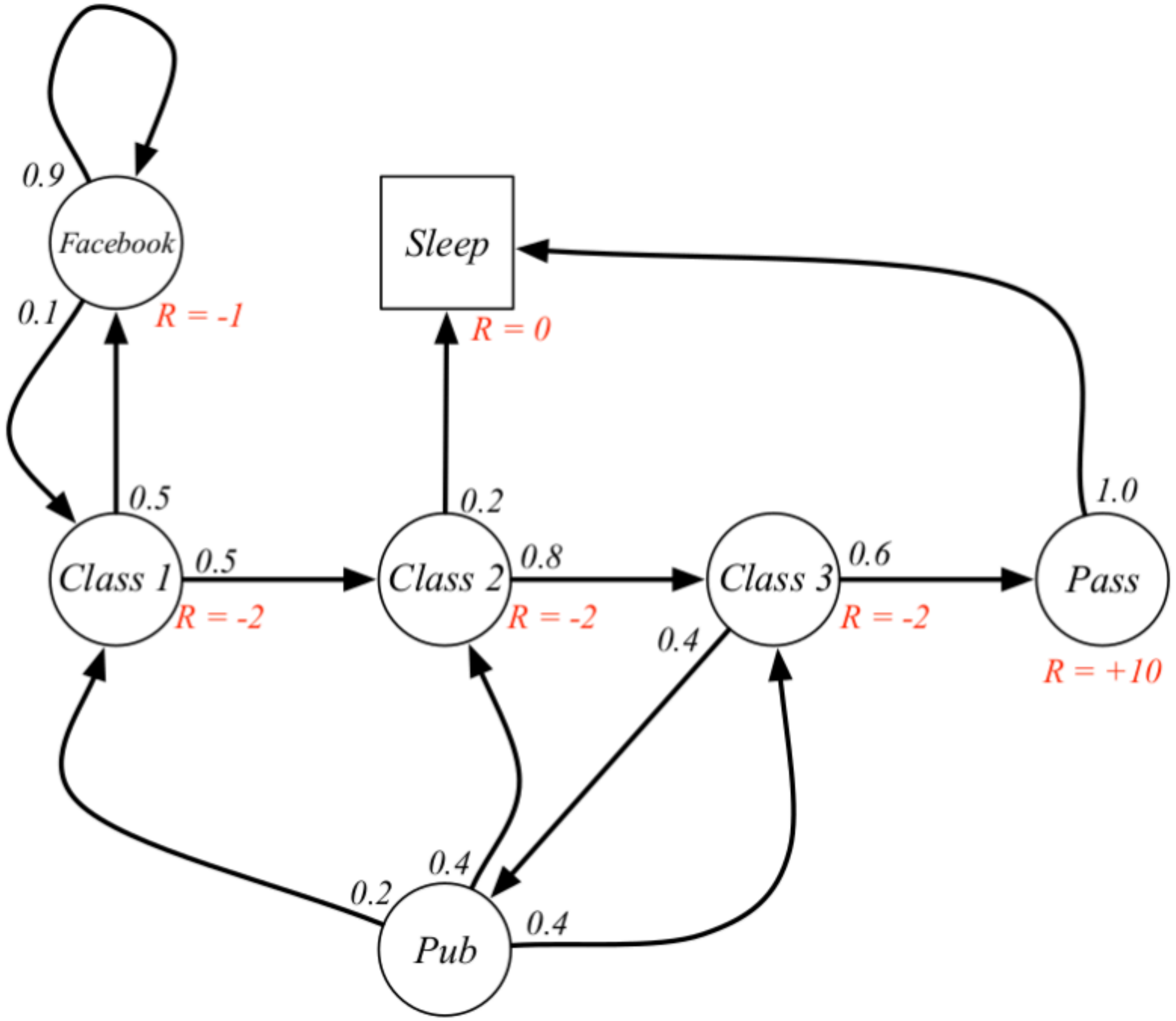
What is the best way to wander?

Readings for today

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

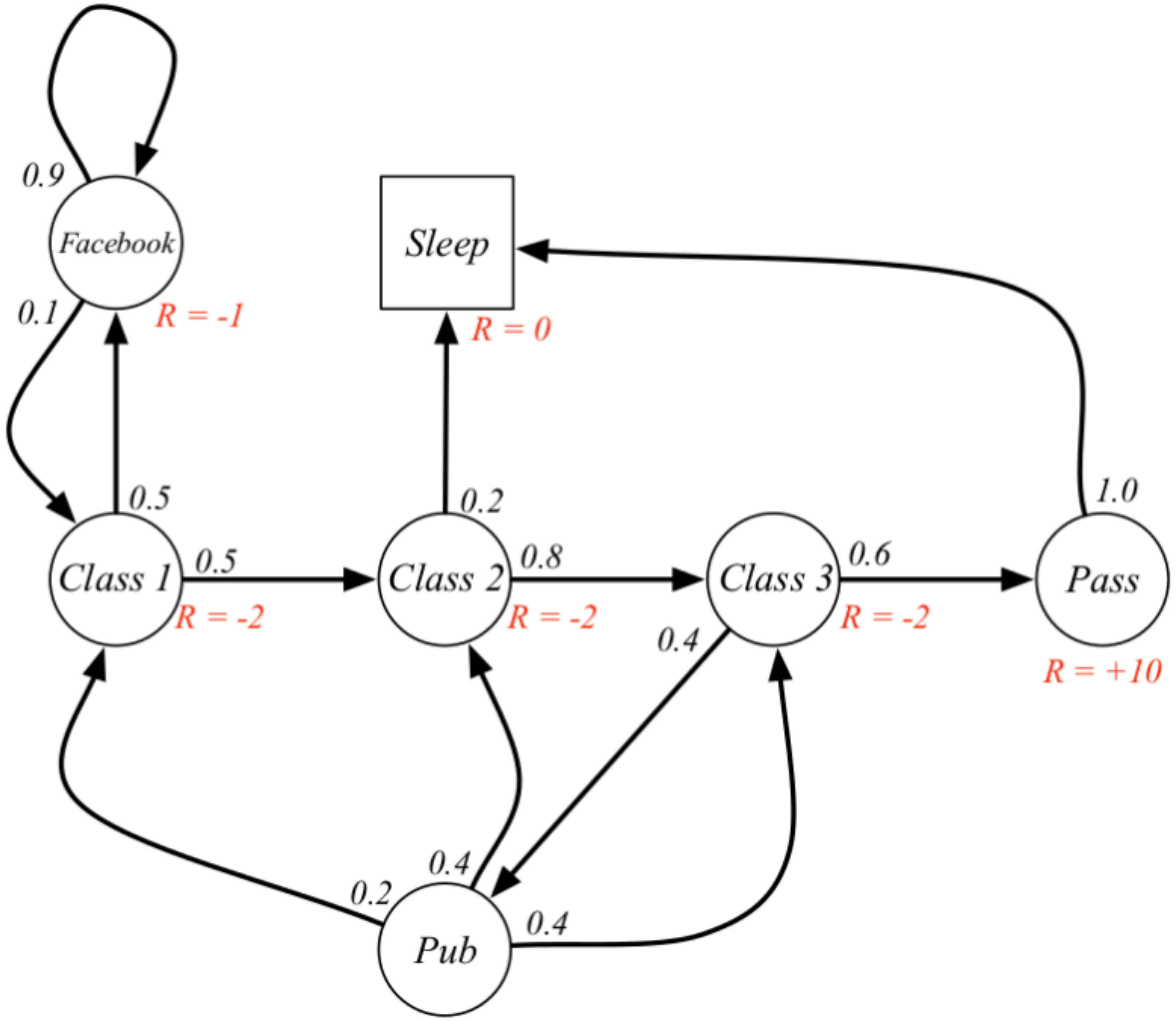
The state-action problem



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

(Ashraf 2021)

Markov property



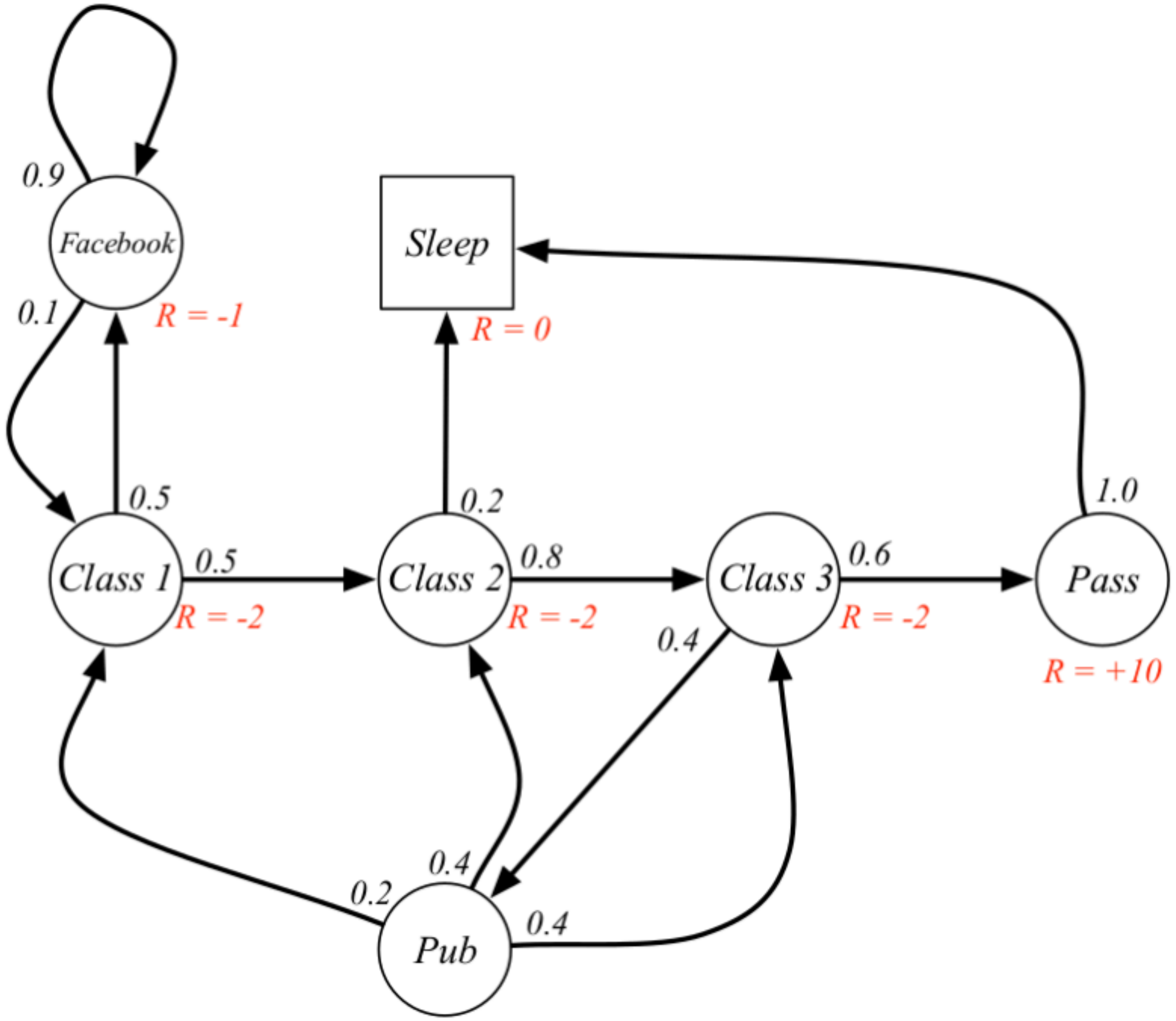
"The future is independent of the past given the present."

State $\mathbf{S}_t$ has the Markov property if and only if

$$P(S_{t+1} \mid S_t) = P(S_{t+1} \mid S_1, \ldots, S_t)$$

(Ashraf 2021)

Markov property

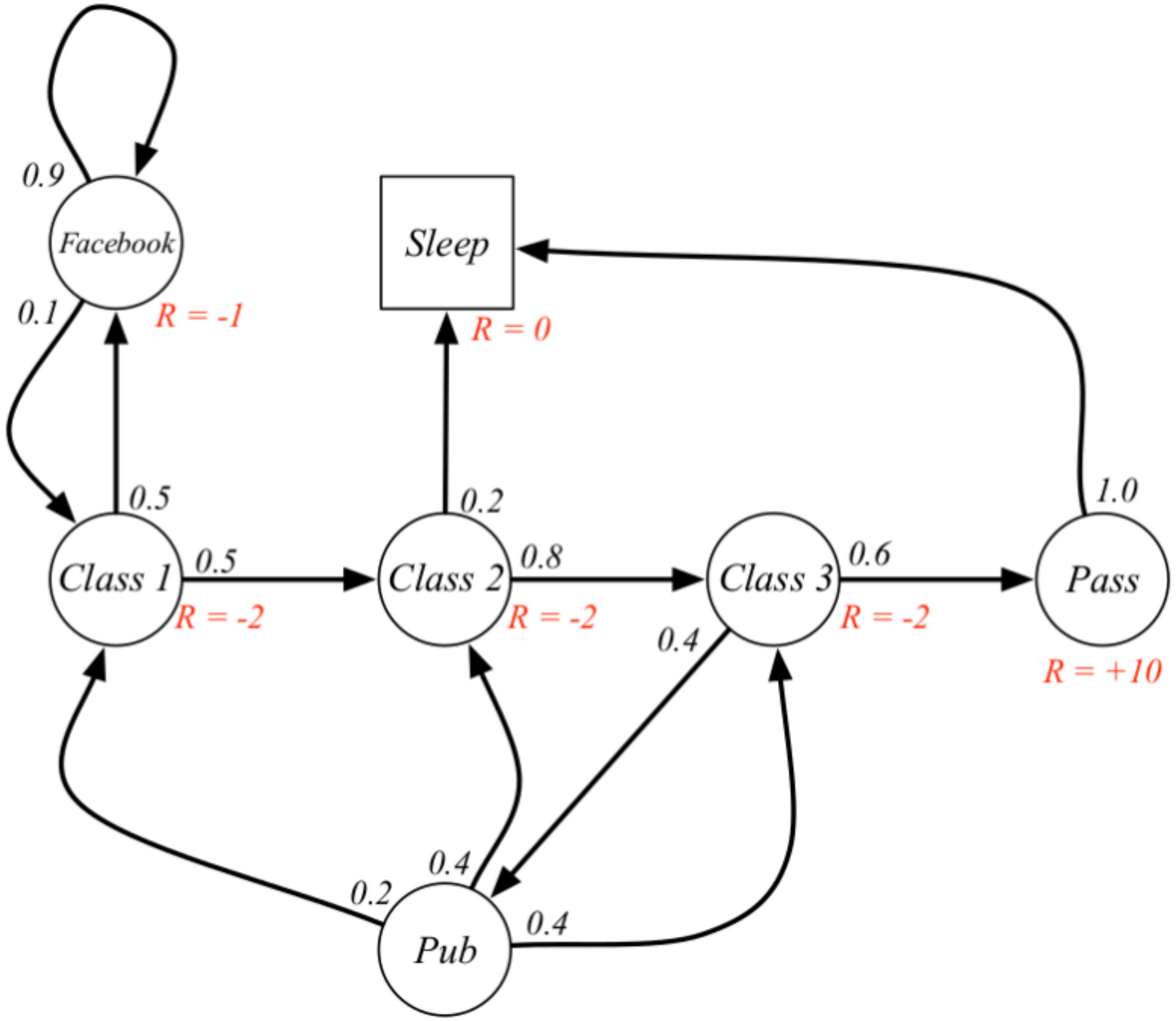


State transition

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

(Ashraf 2021)

Markov process



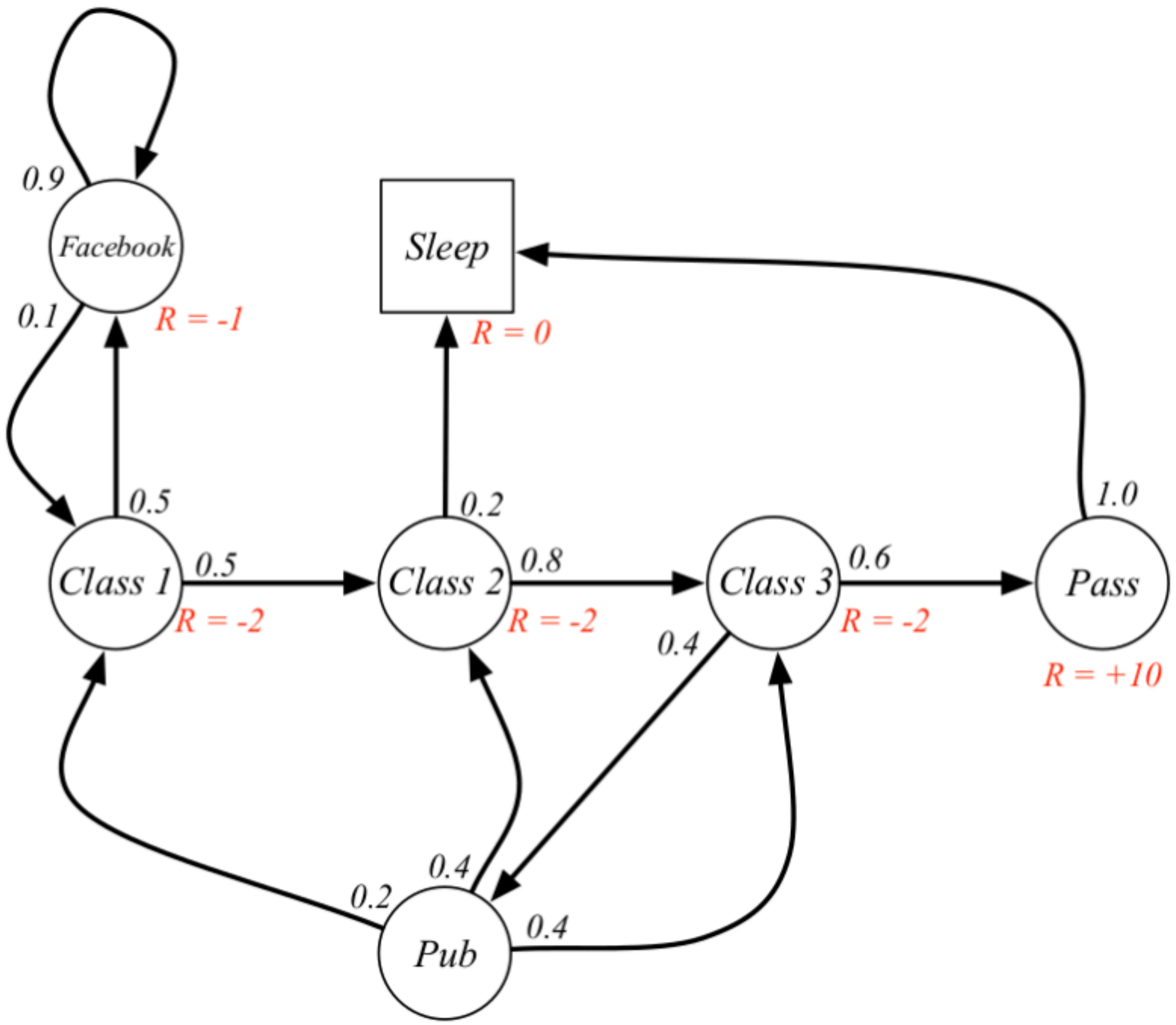
A *Markov process* is a memory-less random process, i.e. a sequence of random states $S_1, S_2, \ldots$ with the Markov property

$$\underbrace{\mathbf{S}}_{\text{state}}, \underbrace{\mathbf{P}}_{\text{transition function}}$$

$$(\mathbf{S}, \mathbf{P})$$

(Ashraf 2021)

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

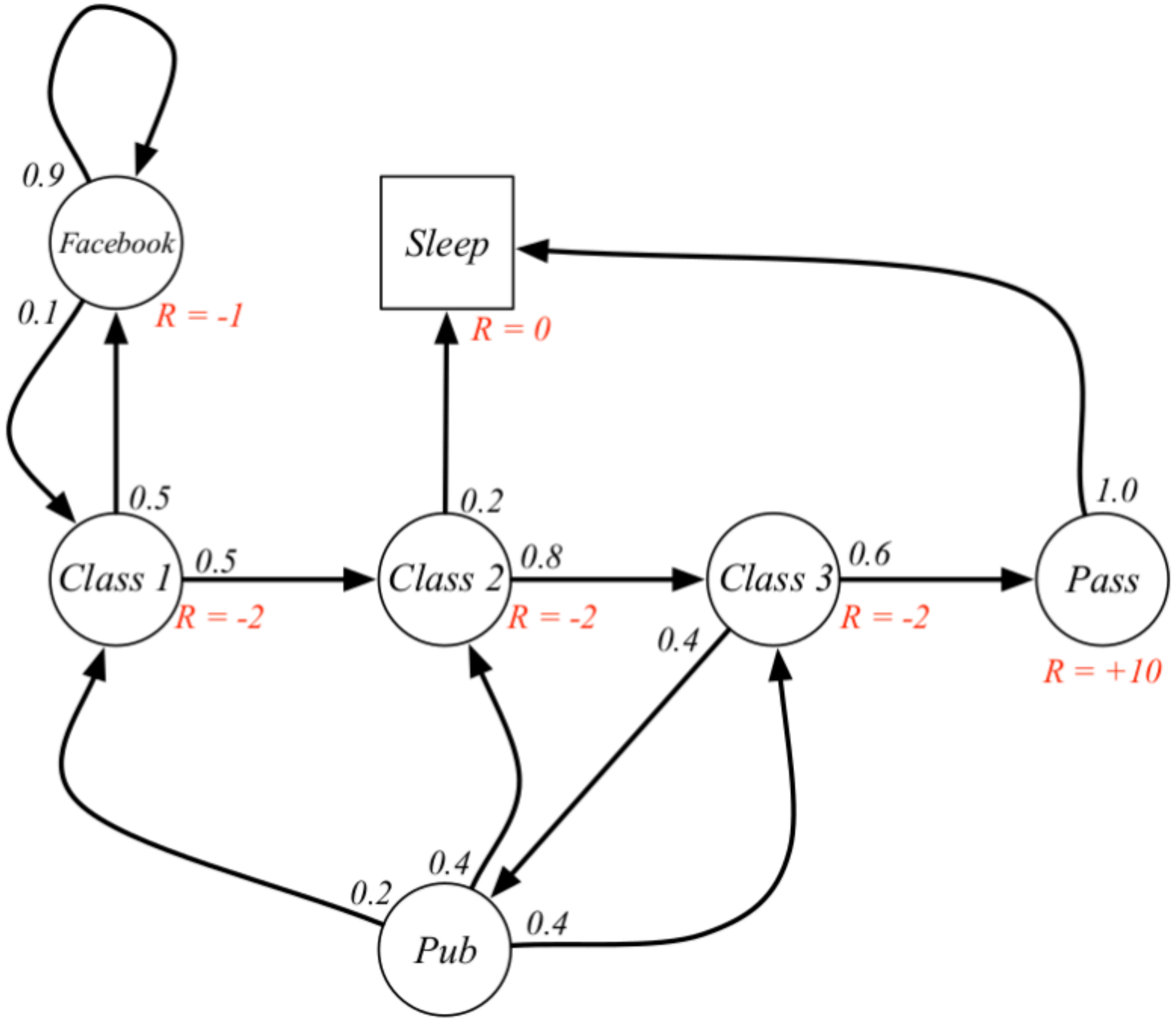
$$(\mathbf{S}, \mathbf{P}, \mathbf{R}, \gamma)$$

state

transition function

(Ashraf 2021)

Markov *reward* process

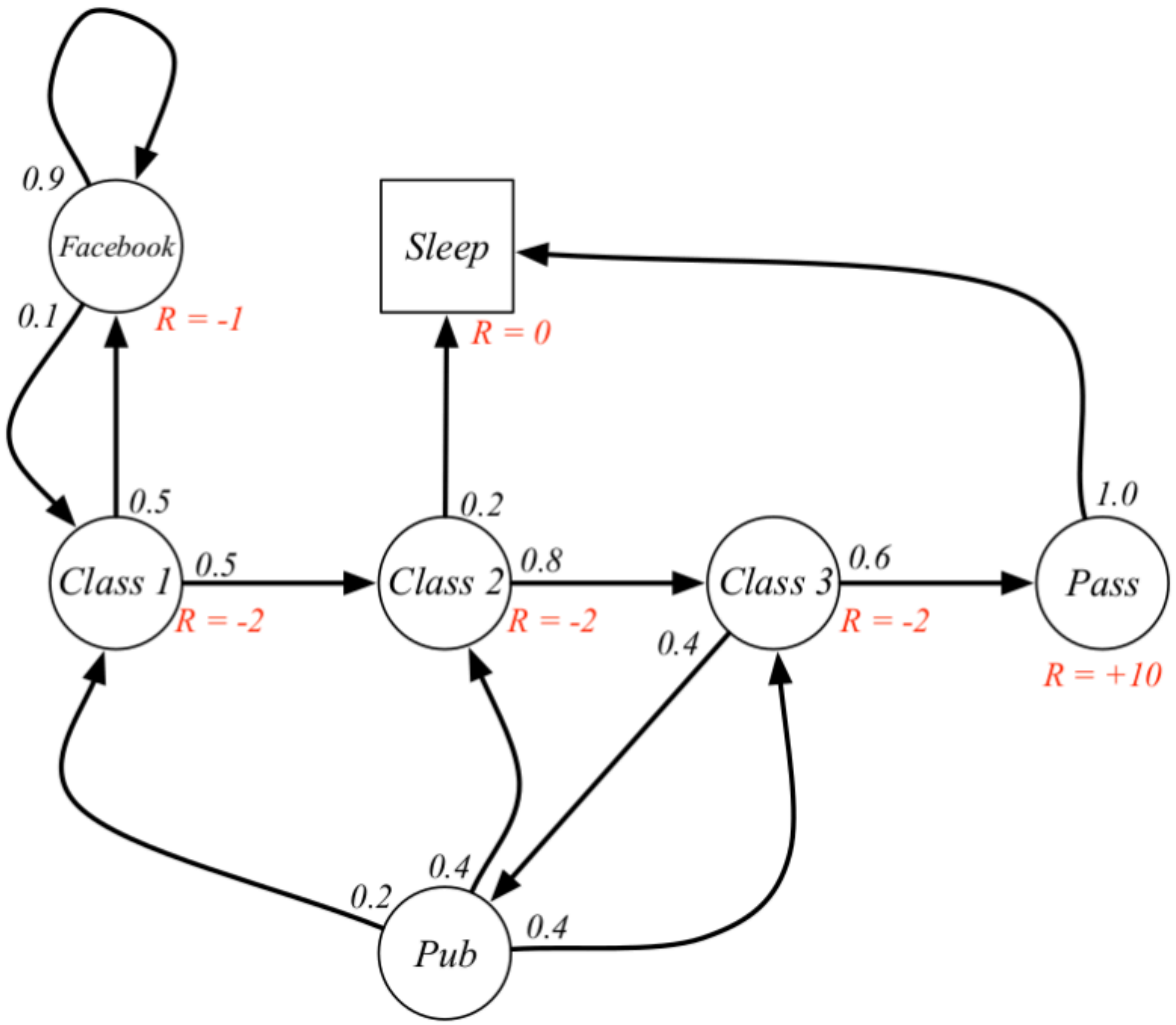


Gain function: Far future awards are less valuable than more immediate awards.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$
$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

(Ashraf 2021)

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 →Class 2 → Class 3 → Pass →Sleep] is:

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

(Ashraf 2021)

The Bellman equation

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

Bellman equation

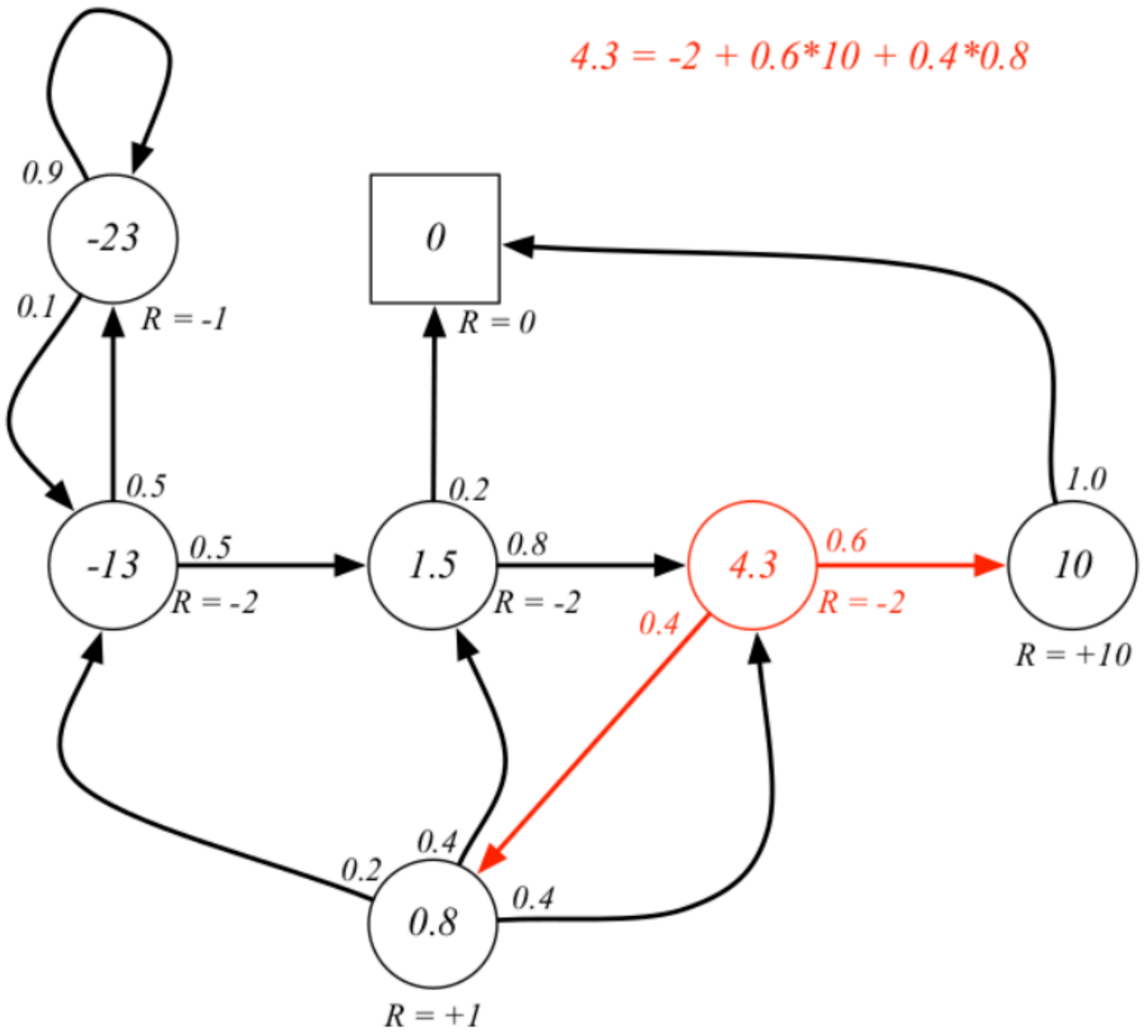
$$\boxed{v(s) = \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) \,|\, S_t = s)}$$

$$
\begin{aligned}
v(s) &= \mathbb{E}(G_t \,|\, S_t = s) \\
&= \mathbb{E}(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \,|\, S_t = s) \\
&= \mathbb{E}(R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \ldots) \,|\, S_t = s) \\
&= \mathbb{E}(R_{t+1} + \gamma G_{t+1} \,|\, S_t = s)
\end{aligned}
$$

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

(Ashraf 2021)

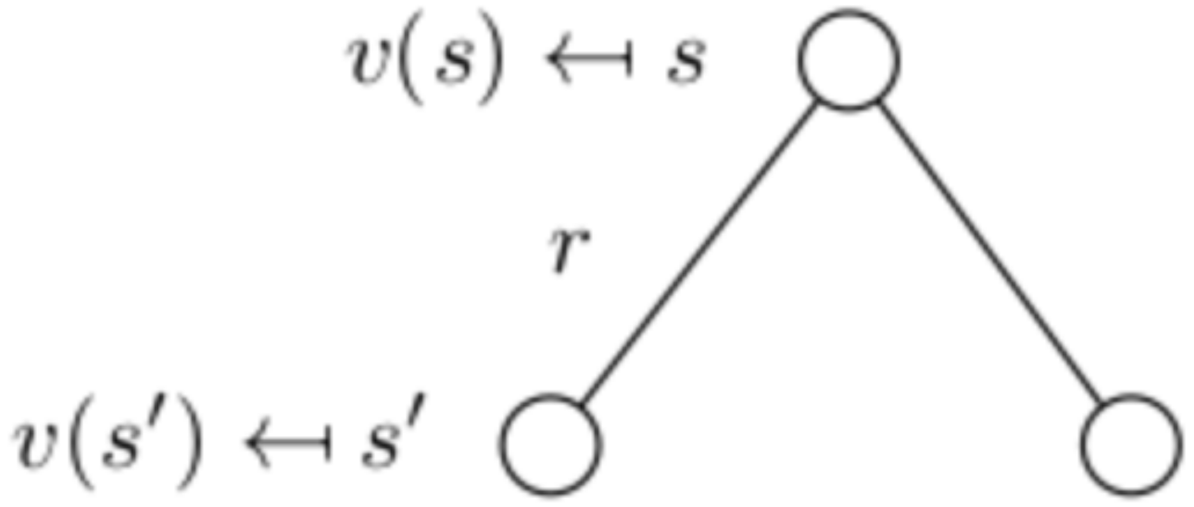
The Bellman equation



The value both depends on the reward <u>and</u> the transition probability .

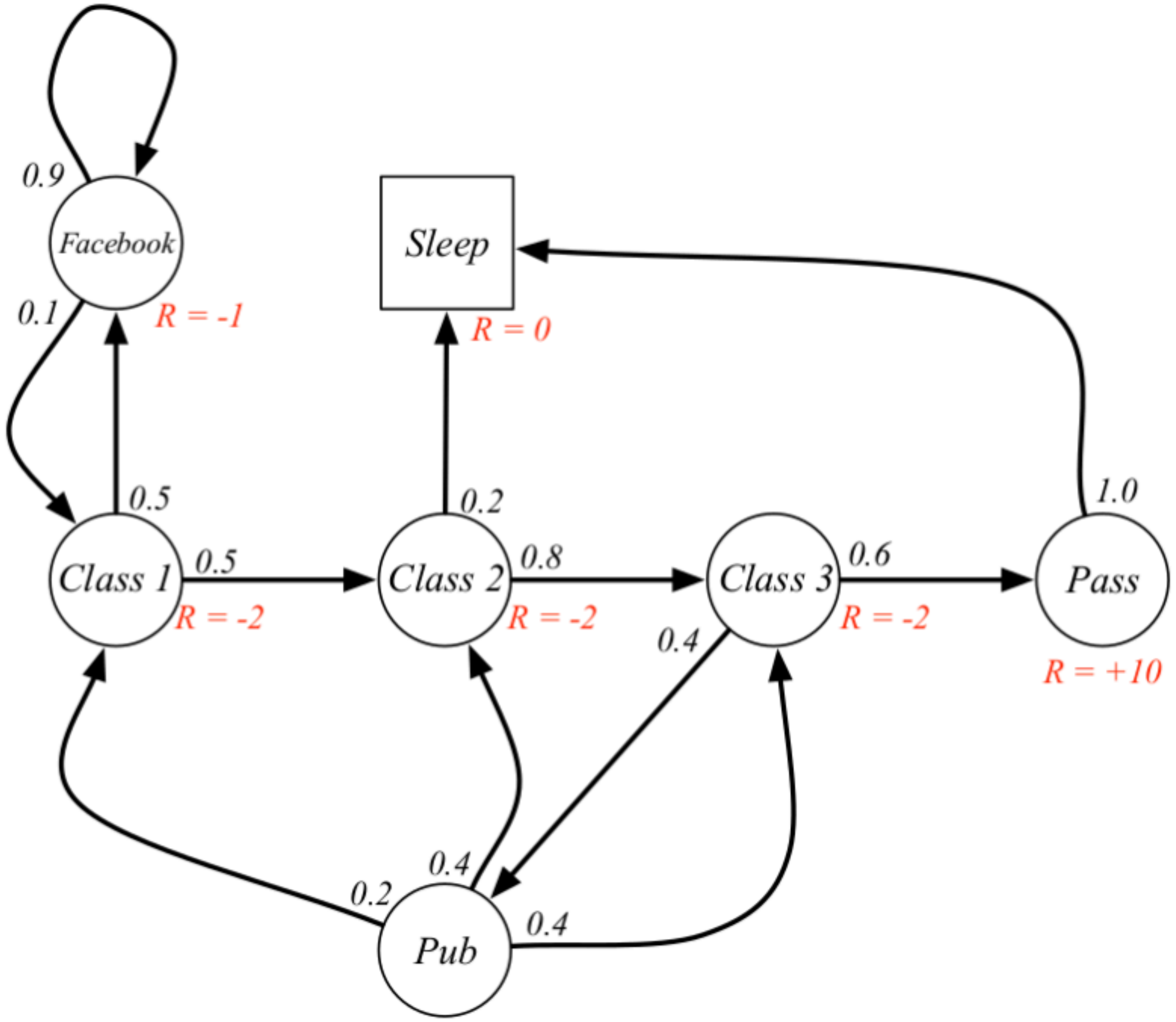
$$v(s) = \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s)$$

$$= \mathbf{R}_S + \gamma \sum_{s' \in S} \mathbf{P}_{ss'} v(s')$$



(Ashraf 2021)

Markov *decision* process



A Markov reward process, with a decision policy π.

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

discount factor

reward function

action

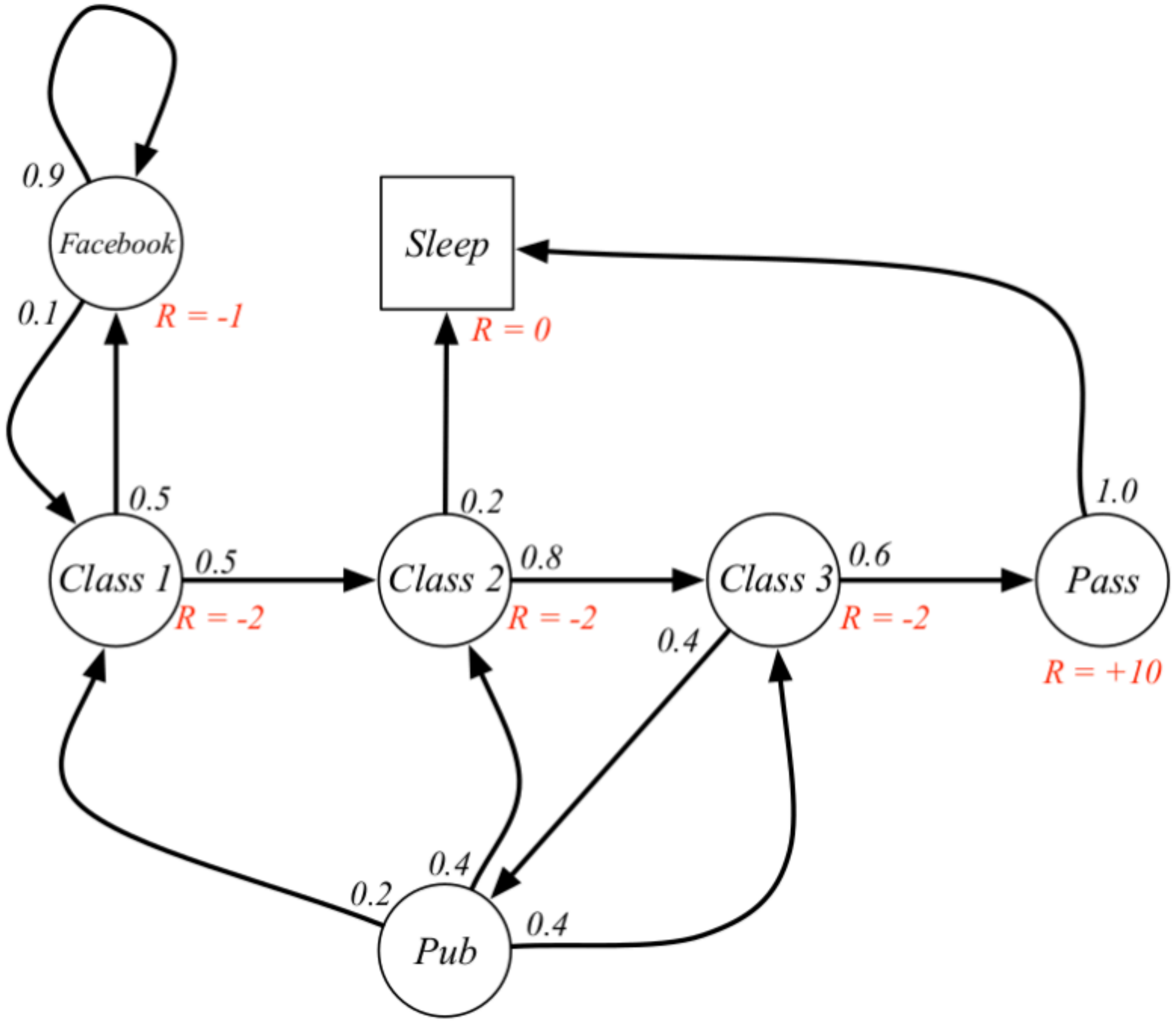
$$(\mathbf{S}, \mathbf{P}, \mathbf{R}, \gamma, \mathbf{A})$$

state space

transition function

(Ashraf 2021)

Markov *decision* process



State transition function

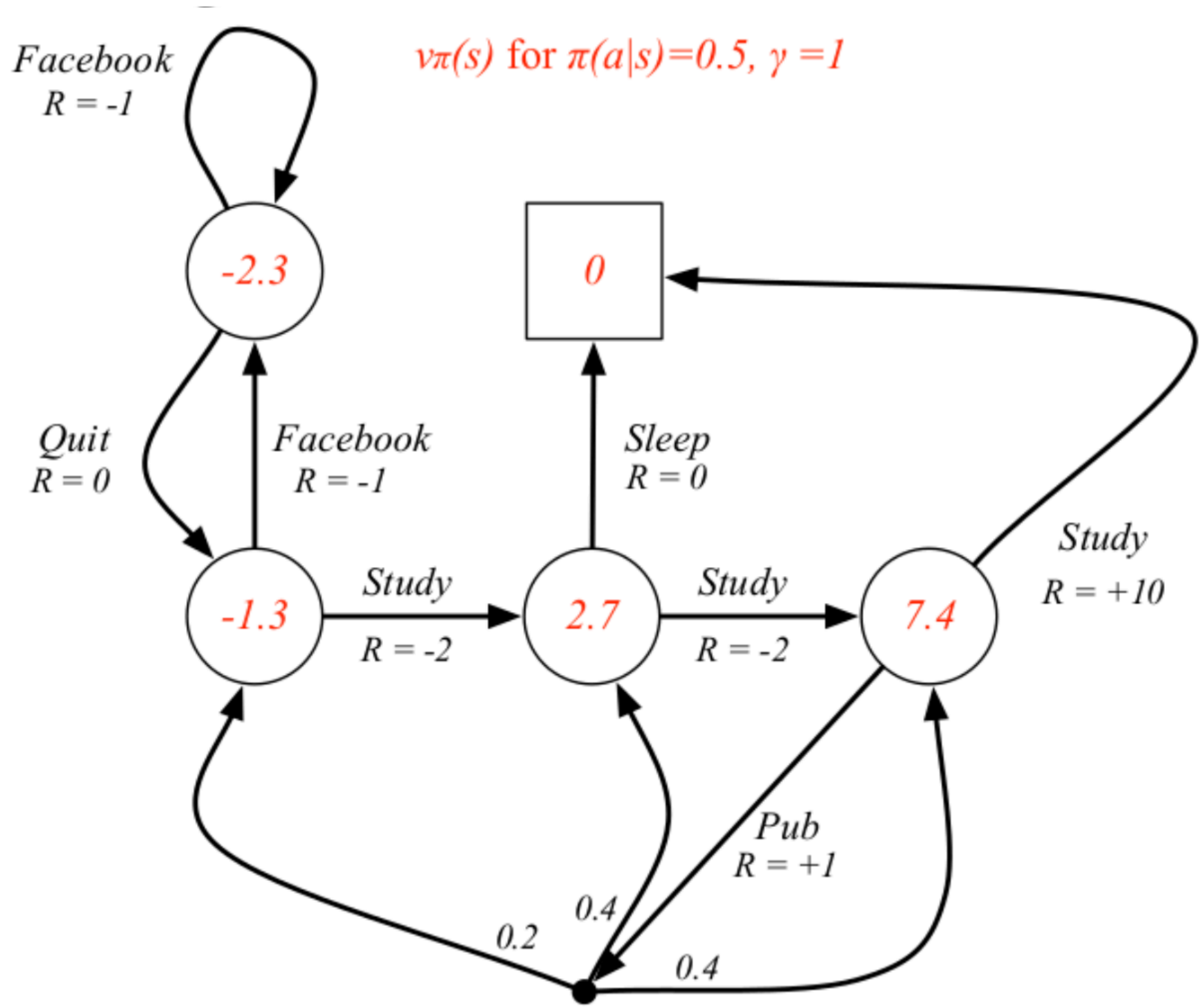
$$P^{\pi}_{s,s'} = \sum_{a \in A} \pi(a \mid s) P^{a}_{s,s'}$$

Reward function

$$R^{\pi}_{s} = \sum_{a \in A} \pi(a \mid s) R^{a}_{s}$$

(Ashraf 2021)

Markov *decision* process

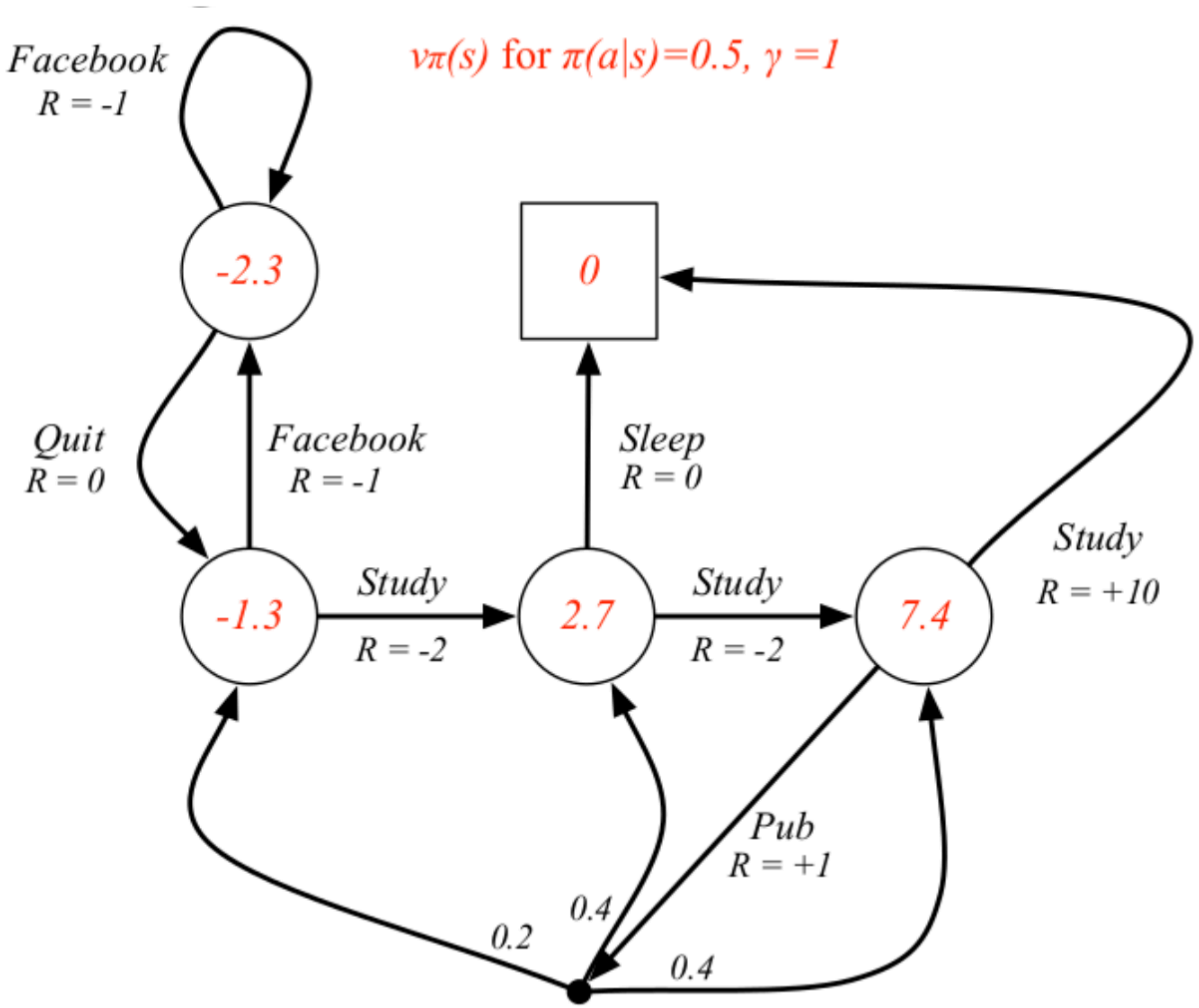


The **state-value** function $\mathbf{V}_{\pi}$(s) of an MDP is the expected return starting from state S, and then following policy π.

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

(Ashraf 2021)

Markov *decision* process



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

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

Take home message

- Markov decision processes (MDPs) show how the future can be independent of the past conditioned on the present.
- MDPs determine the ideal sequence of decisions to maximize *future* rewards when transition probabilities and rewards are known.