Dissociating curiosity-driven exploration algorithms

Interest of curiosity in RL

- Tackling Exploration-Exploitation dilemma.
- Sparse reward problems.
- Emulating human behavior.
- Adaptive and autonomous agents.

Intrinsic motivations

After each transition (s,a,s'), the agent receives an intrinsic reward as:

Novelty: $N^{(t)}(s,a,s') = -\log \boxed{p_N^{(t)}(s')}$ Frequency of s'

Surprise: $S^{(t)}(s,a,s') = -\log \boxed{\hat{P}_{s,a}^{(t)}(s')}$ Estimated probability of transition

Information gain: $I^{(t)}(s,a,s') = KL(\hat{P}_{s,a}^{(t)} || \boxed{\hat{P}_{s,a|s_{t+1}=s'}^{(t+1)}})$ Updated estimation

Empowerment: $E^{(t)}(s,a,s') = \text{Empowerment}^{(t)}(s') = \max_{p(a')} \boxed{I(S'';A'|s')}$ Mutual information

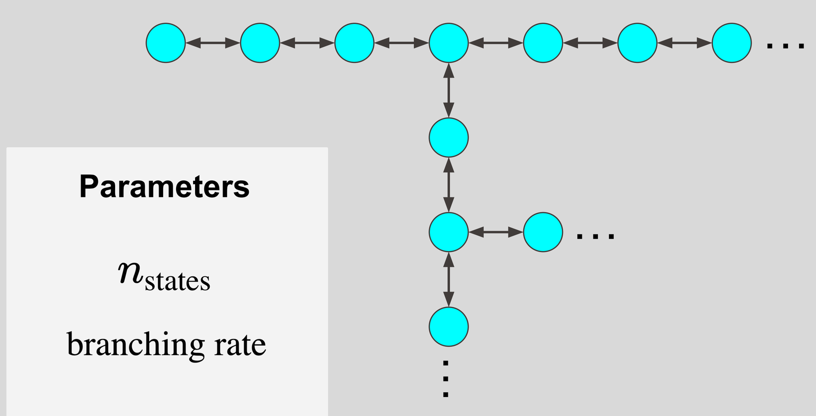
Goals of exploration

1. Visit all the states in the least number of steps.
2. Have a good model of the environment after *n* steps.
3. Visit every state as frequently.

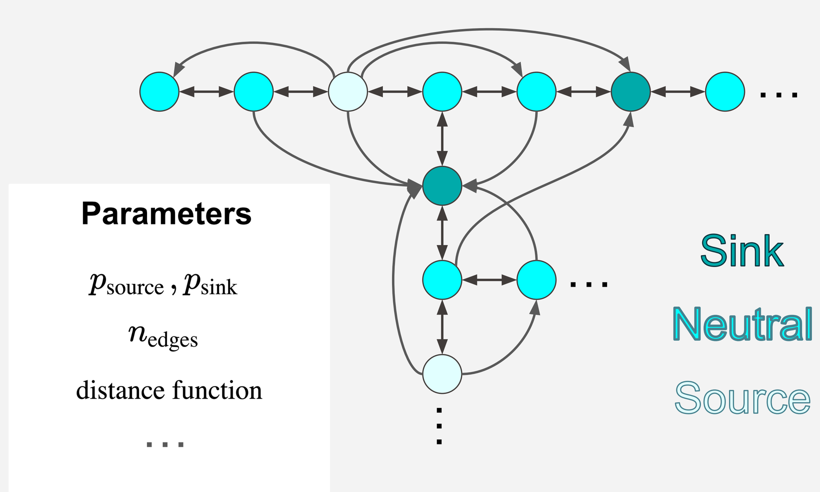
Environment generation

In order to test the agents in diverse scenarios, we design an environment generation process in 3 steps.

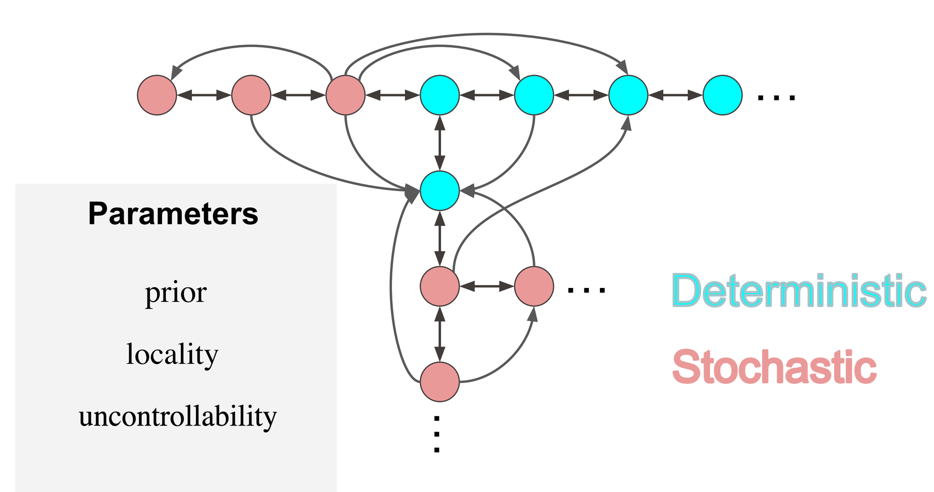
Step 1
Generate a basic structure.



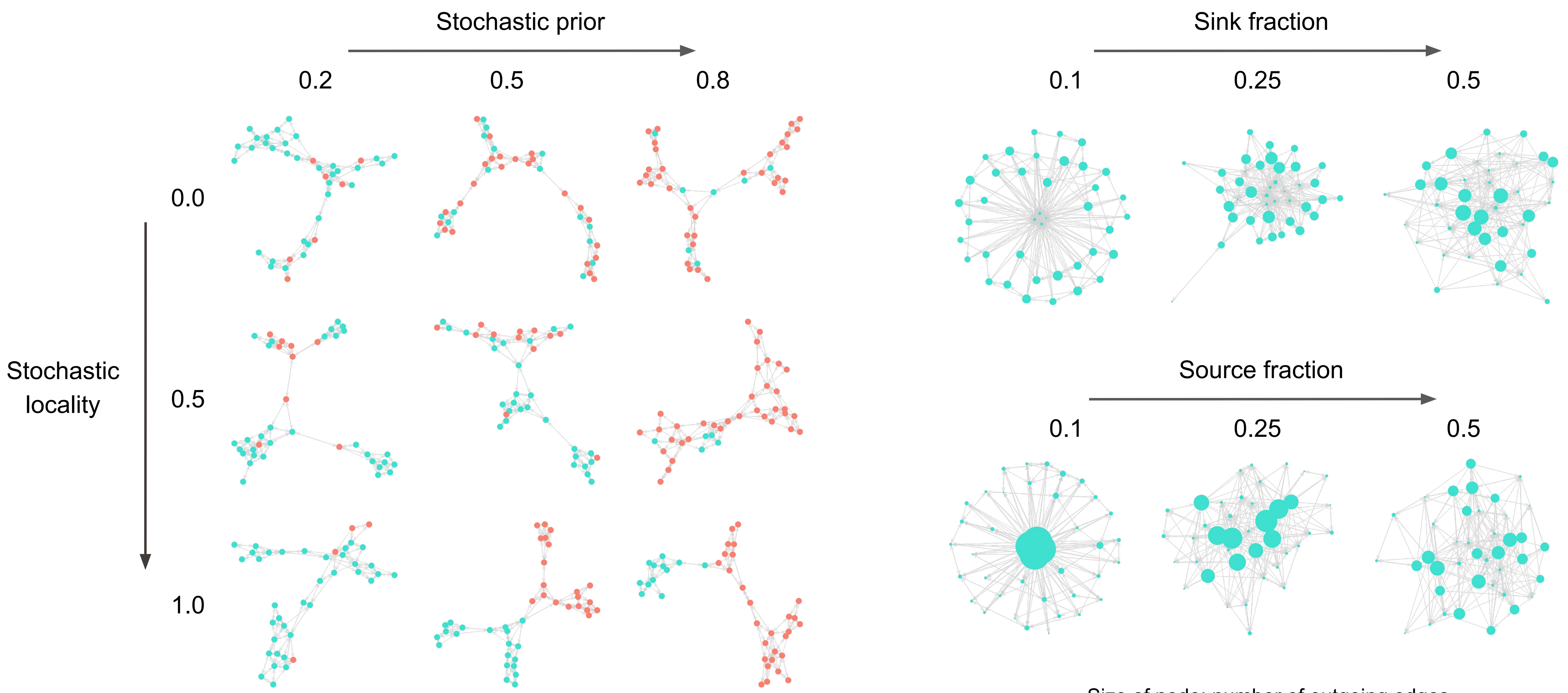
Step 2
Add additional edges, creating sinks and sources.



Step 3
Assign *stochastic* or *deterministic* to each state.



Examples of generated environments, varying parameters.



Size of node: number of outgoing edges.

Results

- Some environment regimes can greatly affect the behavior of the agents (stochasticity, source/sink).
- **Novelty**, **surprise** and **information gain** are about as good to reach all states fast.
- **Information gain** is better to learn a good model of the environment.
- **Surprise** learns fast but stays in stochastic regions afterwards.
- **Novelty** is better for spending a uniform amount of time across the states.
- **Novelty** builds an inaccurate model of the environment when states have numerous actions (sources).
- **Empowerment** is bad for exploring an environment as it tends to stay in empowering regions of the environment.