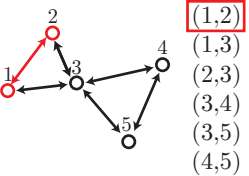
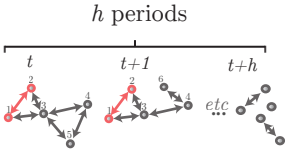
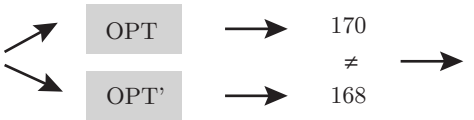


MAB algorithm chooses actions according to its own history of choices and pseudo-rewards.



Environment is simulated until h periods ahead (death times for existing nodes are redrawn)



Optimal matching is computed for each simulation twice. The second time, algorithm is constrained to clear out chosen action immediately.

Matching sizes are compared. Pseudo-reward is 1 if sizes are equal, zero otherwise. Choice and pseudo-reward are concatenated to MAB history. Algorithm repeats until computational budget is over.