1.

a. At the beginning the agent takes a path that leads to a negative reward, then avoids all paths that do not give 0 reward from there on out, because it does not explore and 0 is better than negative.

b. At the beginning the agent takes a few poor reward choices, but by the end the choices are generally very good (at least 100 very consistently). This is because there is a chance for the agent to randomly explore so it learns more about it's environment than just staying on the same 0 reward cycle.

c. (.5) The agent's reward is consistently negative the majority of the time beginning to end. This is because half the time the agent is randomly exploring and not taking the optimal path, so it will usually not be able to take an optimal set of moves.

2. a. The performance of the agent is generally poor, with consistent negative reward values even at toward the end. This is because the agent is not fully aware of its environment (does not know about thief) and has difficulty making proper choices when ignoring the thief's movements.

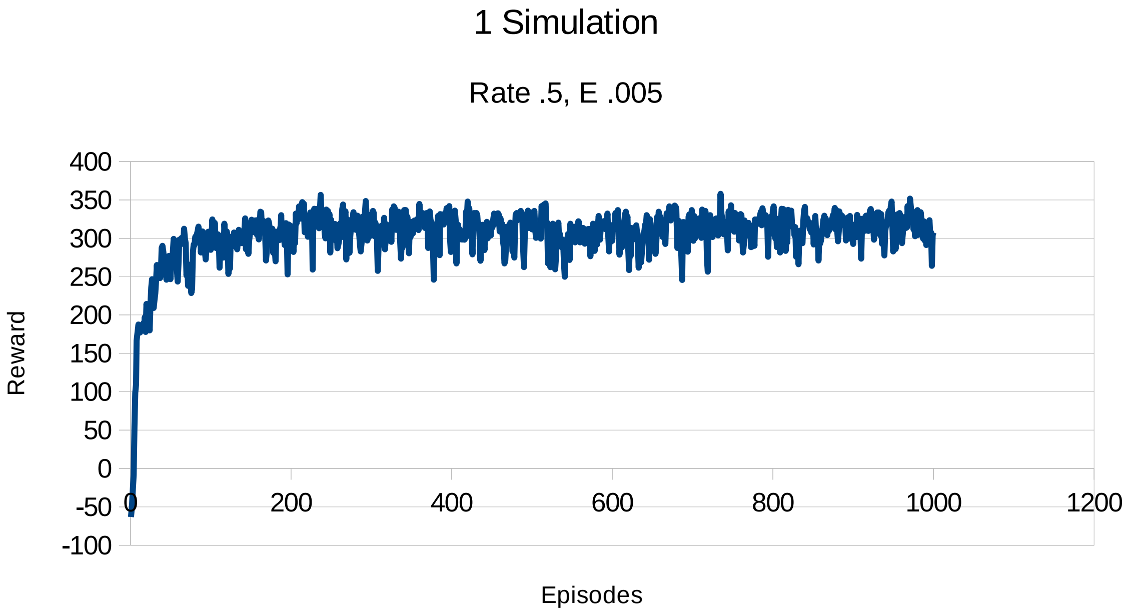
b. The agent initially does poorly, but by the end has learning enough of its environment to get consistently high rewards. This is because now the agent is aware of the thief's existence and can properly learn its environment with that in mind.

c.

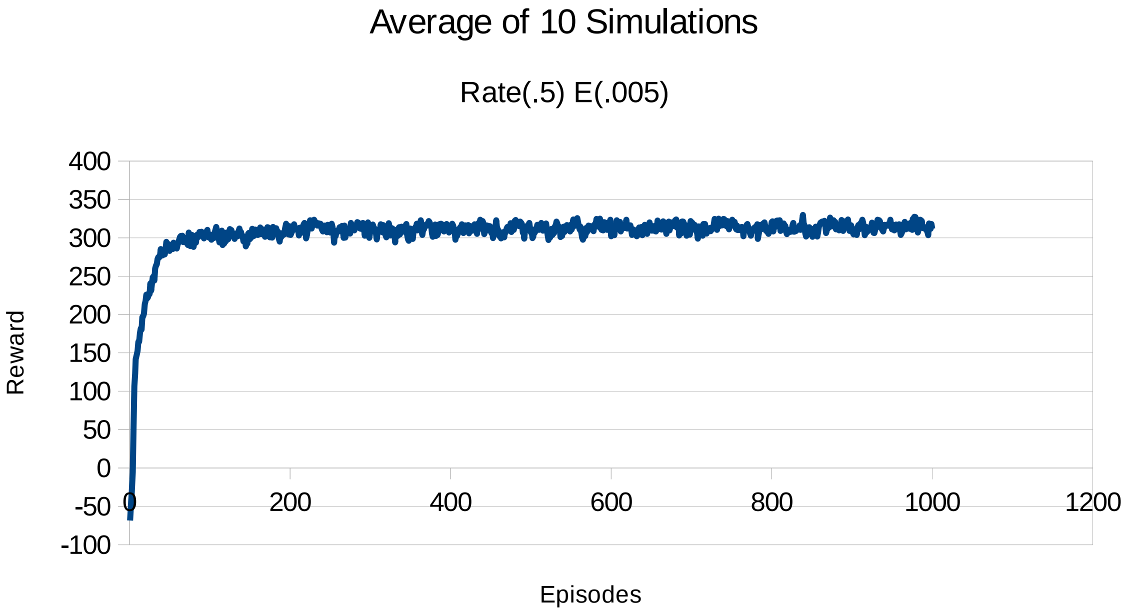
Learning rate- (Tested with e = .05) Anywhere from about .7 to greater than 0. If learning rate is higher than .7, the end rewards are not as high as they could be (less than 200). If the learning rate is 0 then the agent does not do well at all (never learns and acts essentially randomly). However, if the rate is .7 < rate > 0, then the agent can consistently get rewards of more than 200 by the end of the simulation.

Epsilon – (tested with rate = .5) Best e is about .01 to .0005. In that range values at the end are consistently around 290-300. Higher than .01 means that there is too much exploring and it is difficult to take optimal paths, where less than .0005 means the chance to explore is too low so its hard to explore paths not taken yet.

3.



The reward starts off low (less than zero), but rapidly increases to around 175, and continues to increase until fluctuating from 250 to 350.



Now the graph still has a large increase, but the fluctuations are much lower, with values staying stead at around 300 after the initial increase.