15-400 Status Report 1
Andrew Singh
January 28, 2020

1 Major Changes
There have not been any major changes to the project since the last meeting.

2 Accomplishments Since Last Meeting
Since my milestone 1 report from the end of the Fall semester, I have accomplished several things. I have read and written up a summary on a paper about Robust Collaborative Filtering, and I have conducted an additional experiment based on that paper. The purpose of this experiment was to get a better understanding of how adversarial attacks might actually occur in real-world data. Using the attack model described in the paper I read, I simulated adversarial attacks on the MovieLens 1M dataset. I then analyzed the performance of Funk SVD, the collaborative filtering algorithm currently used by MovieLens, on these noisy datasets. The experiment gave me a better idea of how adversarial ratings might manifest in a real-world situation, and greatly contributed to my understanding of the problem.

3 Meeting My Milestone
My planned milestone was to have a solid understanding of the problem, and what causes the model to fail under noisy conditions. I also planned to have started coming up with ideas of how to address the model’s shortcomings. I do have a firm grasp of the problem, and I have a couple of ideas in mind that I plan to explore, so I would say I have met my milestone.

4 Surprises
There have been no surprises to report so far.

5 Looking Ahead
Going forward, I plan to try out some of my ideas to improve model performance under noisy conditions. Firstly, I want to determine if an adversarial attack targeted at a single item is still able to maintain or even increase the model’s performance on the overall dataset. If this is not the case, then one thing we can do is randomly choose subsets of the data and select the subset with the best performance, assuming we know beforehand the proportion of malicious ratings. My other idea to try out was to exploit the similarity between users to identify the malicious profiles. For example, in the “average” attack model, the malicious user rates all the other movies in his chosen set with their average rating, and the target item with the maximum rating. If we can determine that given all of their other ratings, the probability that the user would rate the target item the maximum rating is very low, then we can ignore or deweight that rating.

6 Revisions to Future Milestones
I don’t believe that any revisions to future milestones are needed at this time.
7 Resources Needed

I have access to all of the resources I need for this project.