Yuxuan Zhou (yuxuanz6) – MP1

1.2.3.1 Question 1: Dataset Analysis (10 points)

pb = 0.5; pu1 = 0.96; pu2 = 0.87;
pu1 can't increase anymore, because the unknown locations users with no friends are the minimum isolated points in the graph, and we have no clues to infer them since no points connect to them.
pu2 can be increased, because even if all of their friends don't share home locations, we could use their friends of friends to infer their locations, in that case, we can infer some of their locations and pu2 increased

1.2.3.2 Question 2: Simple Inference (10 points)

The accuracy is: 0.618;
My simple inference algorithm has improved pb from 0.5 to 0.618, that means 11.8% users has been inferred their locations
which is account for 20% of the total unknown location users. It's pretty good.

1.2.3.3 Question 3: Improved Inference Algorithm (20 points)

My improved algorithm takes use of three ideas listed. I don't implement the repeat method which uses the past inferred results.
1. First of all, I use the same way as simple algorithm to calculate geography center point as result 1. Then we start to find the first friend of friend.
2. I find friends of friends who share home locations to the unknown location users whose friends less than 10 in order to get more data for calculation.
3. I then find those shared locations users' points who have not great than 35km from the result 1, get rid of all the ones which is far more than 35 km.
In that case, the users' data could be thorough and concentrated. Thus, the accuracy could be enhanced.

My improved algorithm's accuracy is from 0.618 to 0.631. Although it might not be so higher, but it only repeats for one time. Due to deadline, I don't write repeat method, but it can repeat several times and the accuracy could be enhanced as well.

Performed well:
1. If the inferred method corresponded to the geography center model and the real locations is the center point that we assume
2. The datasets should be large enough, and users should be all kinds of. Their friends should be greater than 10 and smaller than 10, their locations are distributed a lot.

Performed poor:
1. The inferred method doesn't correspond to the center model, there is no clues to predict the real user home locations
2. The datasets are small and concentrated in some direction as a mess