Inference in Location Based Social Networks Report

1. For part 1 of this assignment, I obtained an average accuracy of **61.88697568445822%** for the Within25kmAccuracy test using simple inference algorithm. This simple algorithm does not provide the best results because it is completely based upon the assumption that the user's friends live in the same area as them. As a result, the accuracy is lowered because the users can in fact have friends that live a significant distance away. On top of this, there is a chance that a large number of the user's friends choose to not share their home location limiting the number of points that the algorithm can infer from. So for example, if a user's only friend that chooses to share their home location lives far away, i.e. more than 25 KM, then this algorithm will base its inference off that one location, dramatically lowering the accuracy. So for this particular dataset, the algorithm was not a complete failure and the average accuracy was decent compared to the total number of data points. However, if another data set were used, especially one where the majority of users choose to not share their home location, the accuracy will be significantly lower, which will essentially make this simple algorithm obsolete.

2. For part 2 of this assignment, I choose to take into consideration what the simple algorithm did not. This includes the case where some friends live significantly farther away from the user than the rest of their friends, as well as those users who do not have a lot of friends or a lot of friends that choose to share their home location. In my new algorithm, I first obtain the first friend of a user. I then search for the smallest distance between two friends of the user which I use to assume that these friends live in the relatively same place. Knowing that these two friends live close together, I can use one of them as the initial point for comparison. If a user does not have any friends who share their home locations, I then look at the friends of the user's friend, and find the shortest distance between these friends to use as an initial point for comparison. Once I have my initial starting point, I check if the distance from this point, to any other friends is greater than a certain threshold. In my case, I chose the threshold to be 50 km. If any friend is farther away than 50 km from my initial comparison point, then they are not include in the set of friends that will be used to calculate the user's inferred location. I also check to make sure that the user has enough friends to obtain a reasonable general location. I found that the average number of friends a user has in dataset 1 is approximately 9. So if a user has fewer than 9 friends, I then obtain information from the friends of the user's friends. This allows me to obtain more data points to get a better location inference. This gives me approximately a 10 % increase in accuracy compared to the simple algorithm using dataset 1. It is better than the simple inference algorithm because it eliminates friends of users that are far enough away that it causes the inferred location to considerably inaccurate. It also allows me to obtain more data points for users who would normally not have enough to make a decent inference. I think that my algorithm would not work as well given a dataset where the user only has a few friends, and their friends don't have many friends that share home locations. This would mean less data points to use as well as a higher probability that my initial comparison point may be extremely inaccurate. For example, if a user had friends spread all over the place, the two friends living closest together could be useless if they live in say two completely different states. Other than that, my algorithm will perform especially well when the majority of home locations are shared, or when users with few friends have friends with other friends that share locations in similar places.