Contact networks from trajectories

The goal of this task is to process GPS trajectory data with time information, create contact networks from it and summarize properties of these networks. The task is to be completed in 72 hours (3 days) and solutions should be submitted to *pbogdanov@albany.edu* (see the sections below for details.)

1 From trajectories to networks

1.1 Trajectories and their intersection

A trajectory $T = \{(lat_0, lon_0, t_0), (lat_1, lon_1, t_1) \ldots\}$ is a sequence of position triplets denoting the latitude (lat), longitude (lon) and time (t) of a moving object. This object may be a moving vehicle, pedestrian or any other moving object in 2D. The goal in this part is to identify when two moving objects come into contact, based on co-location threshold parameters $\delta = (\delta_s, \delta_t)$, where the meaning of these two parameters is: the two objects form a contact if they are within δ_s meters from each other at times that are at most δ_t seconds apart.

1.2 Pairwise trajectory intersection

Implement (in a language of your choice) a function that takes as an input two trajectories: T_0 and T_1 and threshold parameters δ and outputs a sequence of contacts in the form of $C = \{(lat_0, lon_0, t_0), (lat_1, lon_1, t_1) \ldots\}$, where each contact point contains the average coordinates and times of the corresponding locations of the two moving objects.

What is the asymptotic $O()$ complexity of your solution if the sizes of the compared trajectories are both n?

1.3 Multiple trajectories from multiple users

Implement a function that takes trajectories from m users (each user may have potentially multiple associated trajectories) and computes contact points from all pairings of trajectories from different users in the form:
$C = \{(u_i, u_j, lat_0, lon_0, t_0), \ldots\}$ (The format is the same as the above, only the user ids are included for each contact).

What would be the overall complexity of this function assuming each user has k trajectories each of size n? Discuss how you arrive to the answer provided.

Can you improve your implementation based on the threshold parameter $\delta = (\delta_s, \delta_t)$? If yes, do not implement it, only discuss and provide pseudo code. Also discuss what would be the change in asymptotic complexity if any (Note: practical improvements may be possible without guaranteed asymptotic changes.)

1.4 A contact network

Assuming a set of contact points C reported in the previous task, construct a contact graph $G(V, E)$ with vertices V the users and edges among users if the two users were in contact at least once. Provide functions to compute the size of the largest connected component in the network and the average degree of nodes. Choose the data structures of your graph representation accordingly.

2 Real-World Data

2.1 Data import

To test your implementation, you will use the Geolife dataset (Download from here: https://www.microsoft.com/en-us/download/details.aspx?id=52367). It contains trajectories for users in China (multiple trajectories per user). Get familiar with the included data file to understand how the data is organized. Write a data parser to be able to load the trajectories in your program.

2.2 Analysis

Consider all trajectories for the first 20 users (folders 000 through 019) and compute the contact network for three settings of the co-location parameters: $\delta_0 = (100m, 300s)$,$\delta_1 = (500m, 600s)$ and $\delta_2 = (1000m, 1200s)$. Compute the corresponding sizes of the largest connected component and average degrees and plot them in two figures, where on the x axis you have the settings of 0, 1 and 2 and on the y axis you have the corresponding values of the largest connected component or average degree. Also record and report the time that it takes to compute the above for each of the settings.

Discuss your observations: Why do you observe the behavior you observe? What do your need to do if you want to increase the average degree? What additional observations can you make?

3 Submission instructions

Submit (1) your code (in a single compressed .gzip file, no executables) as well as (2) the discussion and plots from the previous section in a pdf file to *pbogdanov@albany.edu* no later than 72 hours (3 days) from receiving these instructions. If you have clarifying questions, address them to *pbogdanov@albany.edu*