ASSIGNMENT III (CS 8803 Social Computing)

Due date	3:30pm, November 24
Topic	Analyze and study network and geo-location data from the location based social network Brightkite
What to hand in?	Written report, details based on individual question
How/where to submit?	T-Square
Grade	10% [Total points: 20]

Dataset Source	www.munmund.net/courses/fall2014/notes/LSBNdata.zip
Useful libraries	[Python] NetworkX, SNAP (You are also free to use your favorite programming language, statistical tool, or library)

We had spent a few lectures in September covering analysis of networks arising in social computing systems and problems. This assignment will test your understanding of the concepts and the materials relating to measuring, visualizing, and thinking about social computing system design around social networks.

We will use a dataset from the location based social network Brightkite. Per Wikipedia: "Brightkite was a location-based social networking website. Users were able to "check in" at places by using text messaging or one of the mobile applications and they were able to see who is nearby and who has been there before. The service was created in 2007 […]. In April 2009 Brightkite was acquired by mobile social network Limbo. […] Brightkite allowed registered users to connect with their existing friends and also meet new people based on the places that they go. Once a user "checked in" at a place, they could post notes and photos to a location and other users could comment on those posts."

Questions

Question (1) is compulsory for all. Aside from that, you can do either (2) and (3), or pick (4) and (5). Note each question has different submission requirements.

(1) [10 points] Based on the edge list file of friendship links,

a. Enclose plots on the in-degree distribution, out-degree distribution (*x*-axis will be in or out-degree and *y*-axis will be number of nodes). Logarithmic *x* and *y* axes are better since they visualize the data more clearly. You can use Excel for the plotting purpose, or your favorite graphical software.
b. In a table, report basic statistics of the graph: total number of nodes and number of edges.
c. Compute the following centrality measures on this network: in-degree centrality, out-degree centrality, closeness centrality, betweenness centrality in the table.
d. In the same table, report the number of connected components of the graph, and the mean and median size of a connected component.
e. Include in the table, the mean clustering coefficient of the graph.

You can use the NetworkX Python package for all of the above questions (see [1]), however please feel free to use another library of your choice (e.g., Gephi). Other option: the SNAP library from Stanford, see [2].

What to submit: Two plots for (a) and a table for (b-e).

(2) [6 points] Plot the latitude-longitude pairs for all check-ins in a map. Due to large number of total check-ins, it might be challenging to visualize all points on the map. Hence you should use one of the two methods below.

a. Choose a bounding box size of your choice, so that all latitude-longitude pairs that fall inside it are merged into one group. The color intensity of the latitude-longitude corresponding to the bounding box should be proportional to the size i.e., number of latitude-longitude pairs within that box. The link in [3] contains an appropriate Python script to allow you to do this.
b. Cluster the latitude-longitude pairs first, based on a *k*-means algorithm (see [4]). For *k*-means to work, you will need to compute distances between each pair of latitude-longitude pairs, so for the purpose, you can use the distance measure given in [5], traditionally used to compute distance between two pairs of latitude-longitudes (latitude-longitude pairs measure distances on the surface of a sphere, hence conventional distance metrics like Euclidian distance are not applicable here). After clustering based on a suitable choice of *k*, plot the centroid of each cluster in a map, and vary the intensity of each centroid marker based on the size of the cluster; so if the cluster corresponding to a certain centroid has a large number of latitude-longitude points compared to another that has less, the former centroid should be of a darker shade compared to the latter.

What to submit: Based on the method you choose to plot the latitude-longitude pairs, describe your choice and implementation of the map in one page. It should include, what method you used, why is it a good method for this question, and the libraries you used for implementation. Also enclose a snapshot of the map plot.

OR

(2) [6 points] Propose the design of a web tool or a smartphone application that utilizes your social network structure, your past check-ins, and your friends' check-ins to recommend businesses and things-to-do when you are in a certain area. You need to articulate the method clearly, discuss the design process and the choices of features on the user interface. Submission needs to include a mockup of the tool/application.

What to submit: Two page report on the method, design process, UI elements, novelty aspects of the tool/app and its limitations; additionally enclose a mockup of the tool.

(3) [4 points] Do friends check-in in closeby places? To answer this question, compute a list of "friends" for each node, where you have an edge going from the node to the friend (A→B). Extract the latitude-longitude(s) of check-in of each friend, if any. Then determine the distance of each friend's latitude-longitude pair from the node's check-in latitude-longitude pair(s), if available. Determine the average distance. This way you will get one average distance measure for each node in the graph. Generate a plot of the distribution of these distances, where *x*-axis is the distance (binned), and *y*-axis is the number of nodes with distances in the corresponding bin. Logarithmic *x* and *y* axes are better since they visualize the data more clearly. Ignore a node if the node has no check-ins.

What to submit: A plot of the distance versus node distribution as discussed above.

OR

(3) [4 points] Perform a short literature review on location based social networks. Identify the research questions that have been tackled thus far, and what is outstanding and remains to be done. Identify potential social computing tools that can leverage this rich source of data. Also write a critique on the kinds of privacy and ethical dangers that inadvertent or intentional abuse of such data may pose to users of social computing systems.

What to submit: Two page report on the above items.

References

[1] https://networkx.github.io/documentation/latest/reference/algorithms.centrality.html
[2] http://snap.stanford.edu/snappy/
[3] http://stackoverflow.com/questions/10108368/detecting-geographic-clusters
[4] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
[5] http://www.johndcook.com/python_longitude_latitude.html