File Sharing in an Ad Hoc Network

Data Visualization and Analysis

Final Research Report

Undergraduate Research Associate
Indian Institute of Management Ahmedabad

Student :

Vikramank Singh
Computer Engineering,
VES Institute of Technology, University of Mumbai

Guide :

Dr.Kavitha Ranganathan
Ph.D, University of Chicago
Associate Professor, IIM Ahmedabad

File Sharing in an Ad Hoc Network

Data Visualization and Analysis

Final Report

February 15, 2015

Abstract

An ad hoc network typically refers to any set of networks where all devices have equal status on a network and are free to associate with any other ad hoc network device in link range. We envision a system where any user in the network is equally empowered to generate and distribute Files in any format to the entire network, using his or her mobile phone. The Research mainly focuses on recognition of patterns and clusters in order to determine the traversing pattern of the students in a particular University campus.

The work was basically divided in three parts - Collection of data wherein we tracked the position and travelling movements of a bunch of students using an open source GPS system for a period of 7 days. The IInd part was wherein we were supposed to visualize that data and perform cluster analysis to detect patterns in the movement of the students. The IIIrd part was of simulation of the data over a synthetic model using a simulator.

Previous Work

Previous work included collection of data of the bunch of students which was in a semi-structured way in form of excel sheets *(File 1)*. The data consisted of the TimeStamp, Latitude, Longitude, Speed, Bearing, Accuracy and Elevation of each day for all the students. Each datasheet of a particular student for a particular day consisted of data sets in the range of 5000 - 20000 entries in a semi-structured way. Also the .kml files of each student was generated which could be run on open source software like Google Earth in order to actually visualize the path in which student traversed throughout the campus entire day.

Here onwards the task was to clean the available data and convert it into a structured format and then visualize the data in order to find the clusters or a generic pattern in the traversing of the students and also the sinks i.e those places in the campus which remain crowded for the maximum period of time throughout the day.

Our Work

As explained above, the work started by cleaning the data which included extracting the erroneous data entries also at the same time storing the data in a structured manner*(The structured data is in File 2).* We first converted the Time Stamp which was initially in the UTC format to the IST Time format. Next, we observed that the time interval between every data entry was about 3 to 5 sec and hence we calculated corresponding change in the Latitude and Longitude to determine what as average change in the Latitude and Longitude after every time stamp. Also at the same time we calculated the distance between consecutive pair of Latitude and Longitude for a couple of data entries. Then using these couple of entries making use of Linest Function in Excel, we calculated the corresponding distance each and every consecutive pair of Latitude and Longitude for around 10000 data entries per student per day. Along with the distances (in meters) we also calculated time difference between each timestamp (in sec) which came out to be in the range of 5-13 sec.

Now, using these newly found data entries of Time(in sec) and Distance(in Meters) we calculated the speed of the node (student) for each timestamp in order to determine whether it is Dormant or Moving at a given timestamp and if it is dormant or moving then for how much time.

An average speed of walking for a human is expected to be around 2 m/s[1].We generated Speed (m/s) v/s Time Graph for every student. Following was one of the Graph generated for PGP-I student.

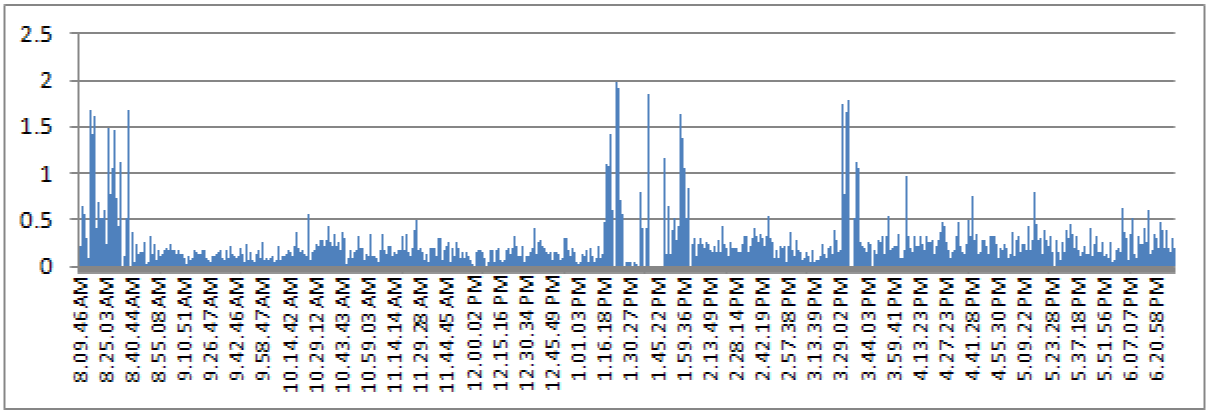

Figure 1: Speed (m/s) v/s Time

The Y-axis represents the Speed in m/s and the X-axis is the Time axis. From the above graph we can see that the for the time period of 8:40 am to 1:20 am the peaks in the graph are too small which indicate that the student was in the classroom (also seen visually using GPS). Similar patterns were generated for both PGP-I and PGP-II students *(Stored in Sheet 1)* .

The Next thing was to detect the clusters. This we began by merging all the Time Stamps which differed from each other with a very small margin. For eg- Time Stamps like 7:08:01am, 7:08:02 am, 7:08:03 am were clubbed together to 7:08 am and similar process was performed for the entire dataset.Their corresponding Latitudes and Longitudes were averaged and the mean value was stored. This was done by extracting the data which was in a structured form in excel sheets to multidimensional arrays using C programming and then complex algorithms were run on the arrays to generate the desired result. Processing all the data by running the algorithms, the resultant data thus obtained was again converted to spreadsheets using files in C.

Similar process was performed on the datasets of all the students for each day. Thus now we had the time stamps in more precise way and their corresponding latitude and longitude. Next step was to determine which student was in the range of the other which was determining the clusters *(Stored in File 3).*

This was done by creating adjacency matrix for all the days where 1 represented the i and j are in range of each other and a 0 indicates they are not. The range was assumed to be 100m. We created an algorithm again in C which calculated the distance among each pair of student for each time stamp and compared that distance value with 100 m and correspondingly entered 0 or 1 in the matrix. The distance between two points was calculated using Haversine Formula[2]

```
dlon = lon2 - lon1

dlat = lat2 - lat1

a = (sin(dlat/2))^2 + cos(lat1) * cos(lat2) * (sin(dlon/2))^2

c = 2 * atan2( sqrt(a), sqrt(1-a) )

d = R * c (where R is the radius of the Earth)
```

Figure 2: Haversine Formula

A typical adjacency matrix is shown below. The timestamp ranged from 7:08 am to 8:55 pm in evening which totalled to 828 data sets.

0	128	72	91	128	134
128	0	72	205	309	250
72	72	0	71	104	88
91	205	71	0	218	139
128	309	104	218	0	379
134	250	88	139	379	0

Figure 3: 6x6 Adjacency Matrix for a particular day of PGP-I

The above matrix indicates that student 1 is in range of student 2 for 128 times out of 828 times in a day. Similarly can be explained for the rest of the datasets. Further expanding this observation, we overlapped the adjacency matrix obtained individually for PGP-I students and PGP-II students to obtain a 12x12 adjacency matrix which represented the movement pattern of all the students together *(Stored in File 4).*

On the basis of the conclusions made from the adjacency matrices, clusters were analysed for both 1st and 2nd year students.
The next step was to come up with a synthetic model on which the data can be simulated to to analyse the theoretical inferences with the one obtained practically i.e after simulation. This we did by identifying the Sinks *(places which were crowded most of the time throughout the day)* and also the time spent by the students at those sinks. Thus, we came up with some of the most crowded places in a campus and also the time spent at those places by the students using the data we had with us. As a result places like Classrooms, Mess / Canteen area, Dorms, Play area, Library emerged to be the most crowded places throughout the day. After this we made analysis of the most crowded pathways and also the less crowded ones. Calculating these things, we generated a reference model of the campus on which we were experimenting which is as shown below.

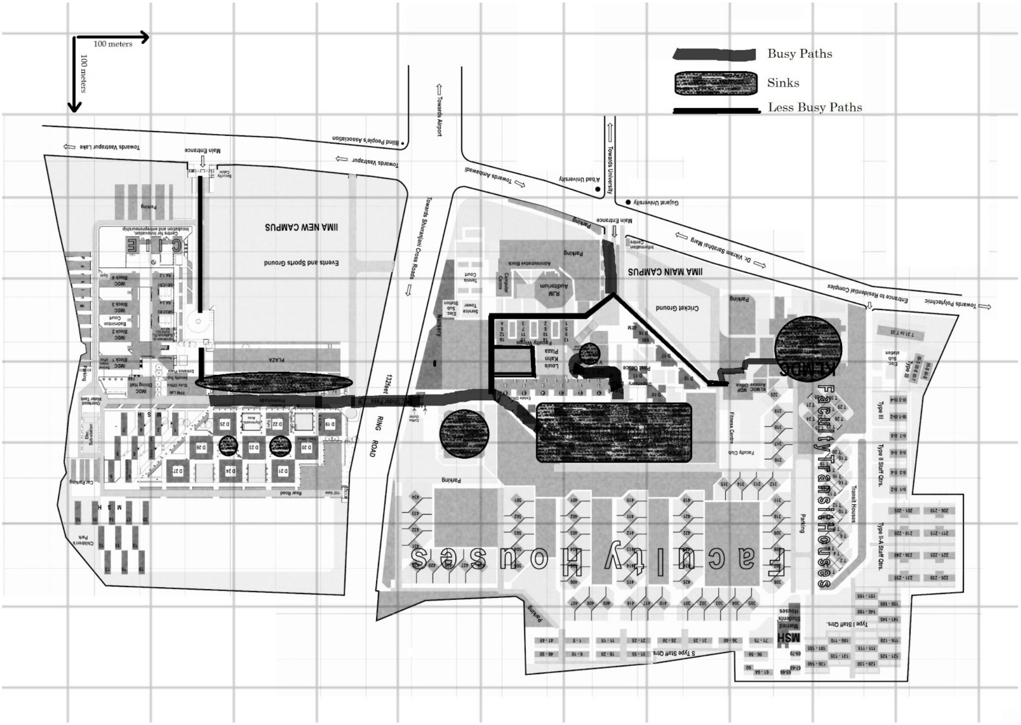

Figure 4: Site Map of a campus depicting Sinks and crowded streets in the campus

Similar observations were done for various institutes in India in order to identify similar patterns and following were some of the obtained patterns.

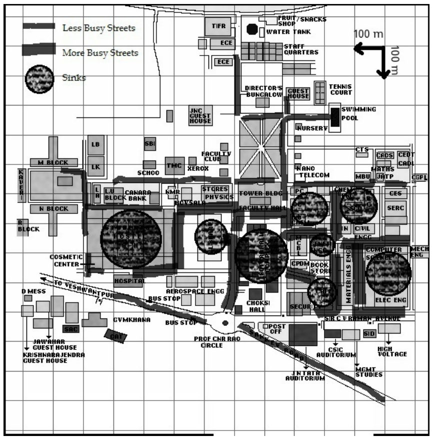

Figure 5

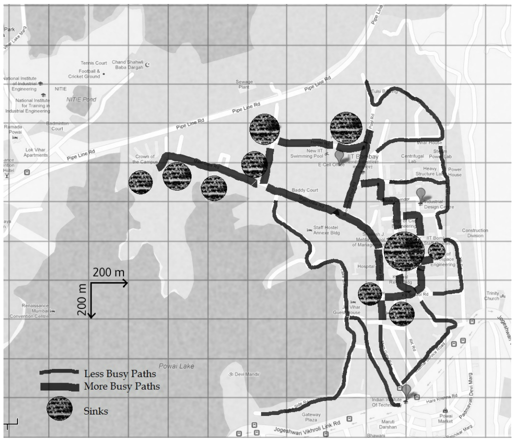

Figure 6

Thus on analyzing the patterns obtained from these sitemaps we get an idea of how the the clusters are distributed on a campus throughout the day and also for what time period *(All the maps are stored in File 5).*

In order to back the results obtained from the visualization of the sinks and paths on various site maps of campuses in India, we also did a data collection of the time spent by the students at these sinks and on the paths using Google Earth *(Data stored in Sheet 2).*

Next Step was to simulate the obtained data on a synthetic model to confirm the formation of clusters among the peers. This was done by creating a grid in an Excel sheet which was then simulated by running VBA Codes on it.

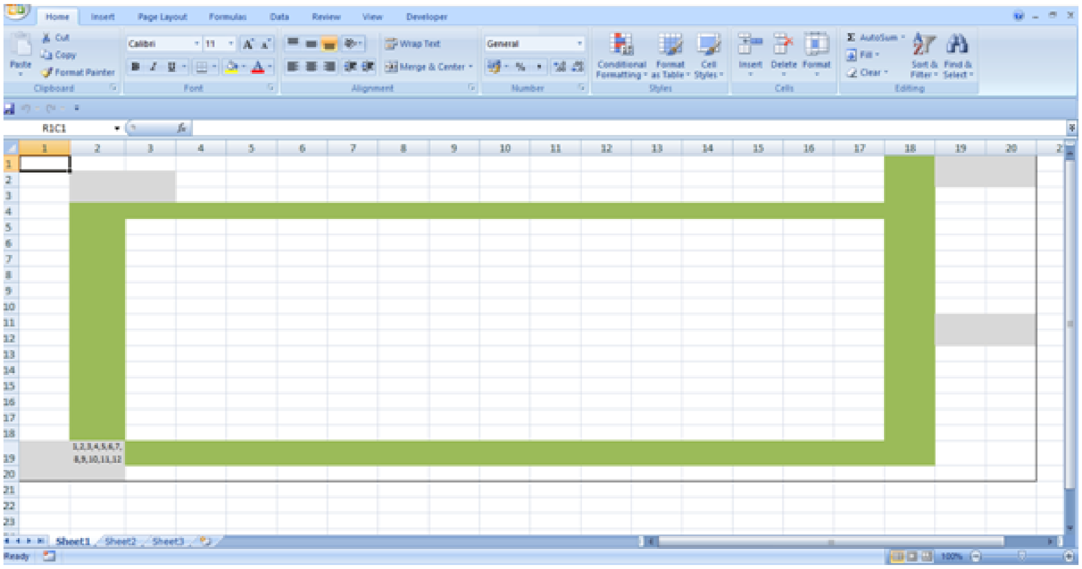

Figure 7 : Grid used for Simulation

The Green patches denote the path in the campus whereas the Grey patches denote the Sinks (Classrooms, Library, Mess). The data in cell (19,2) represent the students where each node represent a student. All the simulations were made on this grid to observe the patterns and clusters. *(All data stored in File 6).*

Future Work

The next step comprises of the simulation of the data on the synthetic model to which would give a practical and more precise results of how exactly the clusters are distributed and would give us an idea of how accurate our theoretical assumptions were.

References

[1] http://www.csie.ntu.edu.tw/~hsinmu/courses/_media/wn_11fall/abr.pdf

[2] http://andrew.hedges.name/experiments/haversine/