CS 6140 Data Mining Project Proposal

***Comparing performances of different clustering techniques on Live Journal dataset.**

(We have revised/changed our initial proposal to this one. Hence sending the proposal too as requested by Professor Jeff Phillips)

1 Team Members

- Padmashree Teeka (u0880562)
- Roshani Nagmote (u0941394)
- Anirudh Narasimhamurthy (u0941400)

2 Data Set

We went through different ideas and decided with the following dataset:
Live Journal
The dataset is of a free online blogging community. This dataset shows the network between those users who are in a group with each other. There can be multiple number of users in a single group.
We are taking this dataset from Stanford site-
`snap.stanford.edu/data/com-LiveJournal.html`

3 Techniques to mine the data

We have chosen following Clustering techniques for the above dataset:

- K Means Clustering
- Hierarchical/Agglomerative Clustering
- Spectral Clustering

4 Interestingness of the problem

We will use clustering techniques for finding communities in the dataset. Communities will consist of user defined groups such that users who have joined a specific group will come under the same community. This information is important for finding the groups of users who are having similar interests. This would help us in predicting to which community a new user would likely be to join.

5 Objective

We are planning to use the above clustering algorithms to find communities in the above dataset and we will compare the performance of the above three clustering techniques.

6 How we obtained the data?

We obtained the LiveJournal social network from Stanford Large Network Dataset collection. It is of undirected type. Live Journal is an online blogging website where users become friends with each other. The data set consists of users as nodes and there exists an edge between those users who are friends with each other and are in a group together.

7 How large is the data?

The dataset consists of about 3997962 nodes and 34681189 edges between the nodes. The text file consisting of this node relation is about 500 MB in size.

8 Format we are storing our data in

We are storing our data in Matrix data type. Sparse Matrix to be specific. There are a large number of users in Live Journal blogging website. Not everyone is friends with everybody else or in the same group with everyone else. But if we want to create a matrix, it will be of size (nodes X nodes) which in our case will be a matrix of the size (3997962 X 3997962) which will be a huge waste of space. Our dataset consists of edges between user nodes who are in groups together. Sparse matrix will set aside space only for those users who are in a group together and not for all the users. This saves us a lot of space and reduces loading time.

9 Processing of original data

We did not have to process the original data into any other format. Since the text file is huge, of 500MB size, we had to split it up into multiple files of smaller size but all in text format itself. The text is in tab-delimited format. This was done only so that we can open the text file in our text editors for analysing purposes.

10 Simulate similar data

To simulate similar data we would basically need to add a node(node no) in the first column and make sure that it maintains consistency and order of nodes. The corresponding row of the next column should/would contain those users who share a relation with the node in the first column. In this way we could simulate our data.

But then again enough care mus taken to make sure the consistency of the realtionships is maintained between the nodes, so that we end up with the correct clusters which would ultimately help us in finding or predicting the communities to which a user might join based on his current data.

If encountered with similar data, we would first need to get it into the same format as our data to proceed with simulation. We would first begin with finding out those nodes which have some relation with each other. For example, if it is social networking data, perhaps we find users who are friends with each other, if it is research papers data, then we find authors who have published a paper together or find those research papers which have been submitted

to the same conference. Once we have this data, we order the first column in terms of nodes and the corresponding row of the next column would be those users or authors who share a relation with the node in the first column. Once we get the new data in this format we can proceed with the same technique as we would with our data set.
As we are using the sparse matrix for this dataset, if we have to apply this technique to a new dataset, it won't cause any problem. It will just create a new sparse matrix and clustering technique will be applied on it.