**Final Project Proposal**

**Goal**

This project will explore the network of stack exchange users for a particular topic as well as analyze the connections between the text of user comments and the score those comments received.

**Data Source Description:**

The data source that I plan to use is the Stack Exchange from Stack Exchange, Inc.,

Stack Exchange is a network of question-and-answer websites for programming like Quora. The most

popular community on Stack Exchange is stack over flow where programmers share questions and answers for coding problem. The data we are going to use in from https://archive.org/download/stackexchange. Which include all topics discussed on Stack Exchange. Each topic includes a XML file which contains data like Posts, Users, Votes, Comments, Post History and Post Links.

**Motivation**:

The reason why I want to use Stack Exchange is that I need to do coding in my work and study and some times I find it’s time consuming to search online. So I am trying to find common features for top rated answers which may help me to refine my searching process in future.

**Plan for doing the work**

I am going to create network from the file. I will use the Posts and Comments files, and extract post id and user id. Users will be the nodes, and edges will be connections between users. I will consider a user connected to another user if they commented/posted on the same post. Then I will run analysis on the network I will compare the top results for each of these to the users with the highest reputation scores in the users file. I also plan to use the Island method to separate into subparts, and then analyze cliques and triads among the user network.

In addition, I am also going to do text analysis and I will use the comments file, specifically the score and text of the comment. I will look for connections between the length of a comment or any particular words in the comment that are associated with a higher score, to see if it is possible to predict the strength of a comment based on text patterns.

**Up front concerns.**

I think mu concerns will be the data preparation since it’s in XML format. Also the data set are very large it may be difficult to visualize the network,