Leading Question

Is the Youtube Community interconnected and what is the longest shortest path from one channel to another?

Dataset Acquisition

Data Format

We have acquired the Youtube data from the 'Stanford Large Dataset Collection,' http://snap.stanford.edu/data/. Each connected component is a group ,and there are 5,000 communities in this dataset. Nodes are Youtube channels and the edges are links to the connected channels in the community. We also plan to use the entire set of data.

Data Correction

The data has been collected from a credible Stanford university website in 2012. If there are any missing entries in the data we will delete them or in the worst case assign the entries as a NULL value.

Data Storage

We will be using the binary search tree as our primary data structure. Our binary tree storage time complexity will be O(h), where h stands for the height of the tree.

Algorithm

The data structure as a whole will be used to analyze the interconnectivity of the Youtube network. The algorithm that will be implemented is Dijkstra's algorithm which has the time complexity of O*(V 2)*O(V^2)*O(V2). And the input to this algorithm will be a youtube channel that will determine how many nodes it will take from the root to the inputted channel. Secondly, we will implement betweenness centrality to find the central channels that are most influential in that community for this dataset using an appropriate algorithm.

Timeline

From the provided data we will create a binary tree that consists of Youtube channels as nodes and the links to them as edges. We will implement a complete BFS algorithm to traverse through the data. In the next week we will implement Dijkstra's algorithm to get the shortest path for the channel. Finally we will research effective betweenness centrality algorithms to analyze the youtube data.