Leading Question

What were the most important papers and who were the most influential contributors to the General Relativity and Quantum Cosmology category in the arXiv archive for scholarly articles from January 1993 to April 2003?

Dataset Acquisition

We have obtained the data from the Stanford list of SNAP datasets:

http://snap.stanford.edu/data/ca-GrQc.html

It is possible to download a text file version of this.

Data Format

As mentioned above, this dataset consists of papers in General Relativity and Quantum Cosmology category in the arXiv archive for scholarly articles from January 1993 to April 2003. The vertices are individual researchers and scholars, and an edge between two researchers means that they collaborated on a paper. These edges are undirected. If a paper was worked on by k people, there would be a fully connected component across those k vertices. There are about 5.2k vertices and about nearly 14.5k edges. We plan to use all this data. The format in the text file is that each line has two numbers separated by what seems to be tab, where each number is an ID for a researcher (numbers are used to preserve the privacy of these scholars).

Data Correction

It is unlikely that there will be many errors, but there are a few possible ones that we could keep an eye out for if needed. For example, if there is a situation where a line does not have two numbers, we should just skip that line. As this data seems to be from a reliable source and was recommended on the 225 page, we do not foresee any other issues we would need to correct.

Data Storage

The data will be stored as one undirected graph. To store vertices and their edges, we will likely follow the same structure as in lab ML (an adjacency list), using its graph and edge classes as a basis for our own. For each line, we would have to at the very least insert an edge. When we encounter a vertex for the first time, we would also have to insert that into the graph. This will likely take linear time, which would be O(n). Unless we are mistaken, creating an edge or inserting a vertex are constant time. The storage is likely also O(n), since the major work will be

in inserting / creating edges for the graph. Also, it would make sense to have the edges be weighted (two researchers might after all work on a paper multiple times for example). This also means that if the edge exists, we would increment its weight by one.

Algorithms

One idea is using a betweenness centrality algorithm to determine the most influential / important researcher in this field in those years. Once we determine that researcher, we can use DFS for our second algorithm / function: find the number of triangles that this researcher is a part of. (Initially we wanted to look at all completely connected subgraphs, but that is unfortunately an NP complete problem, so it is not viable for us). Finally, we will do a graph visualization using force draw algorithms.

Timeline

We are slightly behind, as we have changed our topic. Over fall break, we will try and finish up the data acquisition portion and graph setup sections. Over the next few weeks, we will divide up the algorithms. We should perhaps start with betweenness centrality, as one of our other algorithms depends on the result of that. At the same time, we could probably work on the force draw algorithm, as that would not need us to have the results of any other algorithm. We estimate that the betweenness centrality might take the most time. Once we have that figured out, we can start working on the finding triangles algorithm.