Project Goals

In our project, we will load a dataset of Wikipedia hyperlinks. The dataset contains the names and categories of each article. The format of the dataset is a directed graph with each node containing the name of the Wikipedia article and a value that represents a hyperlink to a different related article. Each article is also categorized which may be overlapping with several other categories. The data is sourced from Wikipedia and was collected in September 2011. In total there are over 2 million hyperlink edges available to us, though we might concatenate the data to make it easier to work with.

Our project will process the data into a graph structure and support a Breadth-First Search (B.F.S.) searching algorithm. Additionally, we plan to implement two algorithms: Cycle Detection- to find the first cycle path from a user defined article ID, and Kosaraju's algorithm- to find the strongly connected components of a directed graph in linear time. If time permits, we also plan on implementing a graph coloring algorithm. Our backup dataset would be an Amazon dataset connecting commonly bought together products.

Source location: http://snap.stanford.edu/data/wiki-topcats.html

Format: Directed, Communities

Backup source location: http://snap.stanford.edu/data/amazon0601.html

Format: Directed