Community detection in massive web graphs

John Karasinski, karasinski@gmail.com

The Internet Archive has been crawling the web for many years (collecting over 10 petabytes of data as of 2012). To reduce this data to a more manageable size, you can consider it as a directed network of nodes (web hosts) that are connected to other nodes by edges (the number of links from one node to another). I have access to an enormous dataset with many nodes and many trillions of links.

There are a number of interesting things that could be analyzed in this dataset. From a practical side, the Archive is especially interested in:

- Community detection
- Detecting link farms

Though the Archive is attempting to crawl the entire web, link farms are not particularly interesting, and they'd love to ignore them. Detecting these link farms could save them significant processing power. From a more theoretical side, the computational power required to analyze these enormous graphs may not be trivial. There are a number of algorithms designed for analyzing massive graphs. I'm interested in using techniques such as the Louvain Method to extract the community structure of the network. It might be interesting to use different algorithms on the data and compare the results.

I'll be travelling down to visit the Archive on Friday (4/22), and anyone who's interested in either the Archive or this project is welcome to join me.