Web Graph Analysis on the Blue Waters Supercomputer

George M. Slota (gms5016@psu.edu, www.gmslota.com), Penn State University
Kamesh Madduri, Penn State (advisor), Siva Rajamanickam, Sandia National Labs (advisor)

SAND2015-6705C

Research Motivation

Why should we analyze and attempt to understand the structure of the Internet?

- Classifying web site (spam)
- Identifying possible cyberattack vectors
- Understanding the social and economic influences that drive its growth

What makes supercomputers ideal for web graph analysis?

- Web graphs are large and complex (3.5 B+ pages and 129 B+ hyperlinks)
- Supercomputers are more energy efficient with faster turnaround time than approaches that use external memory (MapReduce/Hadoop, flash, etc.)

Contributions

We implement several algorithms for large-scale graph analyis and use these to analyze the 2012 Web Data Commons hyperlink graph on the Blue Waters supercomputer:

- Weakly Connected Components
- Strongly Connected Components
- Approximate K-core
- PageRank
- Harmonic Centrality
- Community Detection

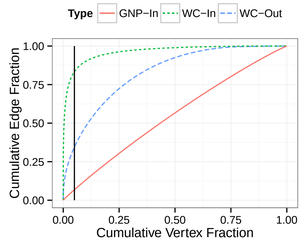
We are the first group to our knowledge to analyze this graph on a distributed-memory system. We also reveal previously unknown insights into the community structure, page centrality, and general characteristics of the web graph.

Downloads

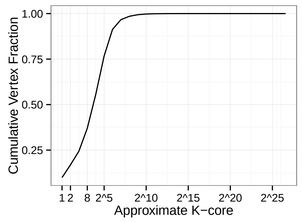
Below we include links to our source code, the full results, and the hyperlink graph:

Source code | Analysis Results | WDC Graph

Global Characteristics and Computational Challenges



In-degree is very skewed: 5% of vertices own 83% of all incoming edges. **This skew imposes challenges for scalable graph algorithm implementations due to potential work imbalances.**



We calculate approximate K-core value and note they are quite large; 20 M vertices have a value of over 1 K. **This imposes additional challenges for balanced work distribution on a distributed system.**

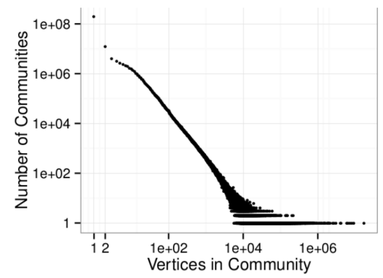
Community Identification

n_{in}	m_{in}	m_{cut}	Representative vertex
112	2126	32	www.youtube.com
18	548	277	www.tumblr.com
9	516	84	creativecommons.org/..
8	186	85	wordpress.org/..
7	57	83	www.amazon.com
6	41	21	www.flickr.com/..
6	39	58	askville.amazon.com
4	133	142	www.google.com
4	280	18	tripadvisor.com
3	78	13	www.househunt.com

Communities in Web Crawl (numbers in millions)

- We plot community size frequencies
- The plot shows similarity to frequency plots of in- and out-degree, WCC sizes, and SCC sizes in Meusel et al. 2014
- This lends additional credibility to the **instrinsic power law but heavy-tailed structural characteristics of the Internet**

- 10 largest "website communities" by number of contained pages (n_{in}) output by our algorithm
- Observing the number of intra- (m_{in}) and inter- (m_{out}) community links, **top communities are dense**, with a high intra- to inter-link ratio
- The communities output by label propagation (Raghavan et al. 2007) are very stable, **identified communities are static between runs and after about ten iterations**



Centrality Measures

Out-degree	In-degree	PageRank	Harmonic
photoshare.ru/..	www.youtube.com	www.youtube.com	wordpress.org
dvderotik.com/..	wordpress.org	www.youtube.com/t/..	twitter.com
www.zoover.be/..	www.youtube.com/t/..	www.youtube.com/testtube	twitter.com/privacy
cran.r-project.org/..	www.youtube.com/..	www.youtube.com/t/..	twitter.com/about
cran.rakanu.com/..	www.youtube.com/t/..	www.youtube.com/t/..	twitter.com/tos
www.linkagogo.com/..	www.youtube.com/..	www.tumblr.com	twitter.com/account/..
www.cran.r-project.org/..	www.youtube.com/t/..	www.google.com/intl/en/..	twitter.com/account/..
www.fussballdaten.de/..	gmpg.org/xfn/11	wordpress.org	twitter.com/about/resources
www.fussballdaten.de/..	www.google.com	www.google.com/intl/..	twitter.com/login
www.fussballdaten.de/..	www.google.com/intl/..	www.google.com	twitter.com/about/contact

- The above table shows the results for various centrality measures, including in-degree, out-degree, PageRank, and Harmonic Centrality for the full page-level web graph
- Out-degree is observed to be relatively meaningless as a centrality measure; however, this might be an artifact of crawl methodology
- We note similarities between the highest-ranking vertices by centrality and the communities shown above as well as prior analysis of the host-level graph in Meusel et al.

Implementation Considerations

There were several factors and considerations we used to get high performance for end-to-end (I/O-preprocessing-execution) analysis of the web crawl.

- We striped the 1 TB graph file across 160 storage units of Blue Waters' scratch filesystem and ingested the file in parallel
- Due to inherent locality within the crawl, we were able to easily create a quality 1D $\frac{n}{p}$ partitioning with an efficient CSR-like per-task graph representation in memory
- Used MPI + OpenMP model with one multi-threaded task per node
- Implemented hierarchical queues to minimize data movement (thread-task-global)
- When possible, exploited communication and computation avoidance strategies

Performance

Alg.	Web Crawl Total	Web Crawl % Comm.	$G(n,p)$ Total	$G(n,p)$ % Comm.
I/O	47	-	0	-
PrPr	150	73%	67	45%
WCC	131	34%	84	12%
SCC	148	25%	88	15%
AKC	1374	30%	731	12%
PR	930	91%	1210	30%
HC	92	34%	41	12%
LP	1438	70%	2406	23%

- Above is the execution time in seconds of I/O, preprocessing, and each algorithm along with portion of total time spent in communication for running on 4 K cores of Blue Waters with the web crawl and equal-sized $G(n,p)$ graph
- I/O and preprocessing completes in minutes for both of the 129 Billion edge graphs
- Algorithms all complete in minutes; about 3× faster than fastest known alternative approach (Zheng et al. 2015)
- Times are often lower for the random graph due to less degree skew and therefore better work and comm. balance among tasks

Acknowledgements

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070, ACI-1238993, and ACI-1444747) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This work is also supported by NSF grants ACI-1253881, CCF-1439057, and the DOE Office of Science through the FASTMath SciDAC Institute. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.