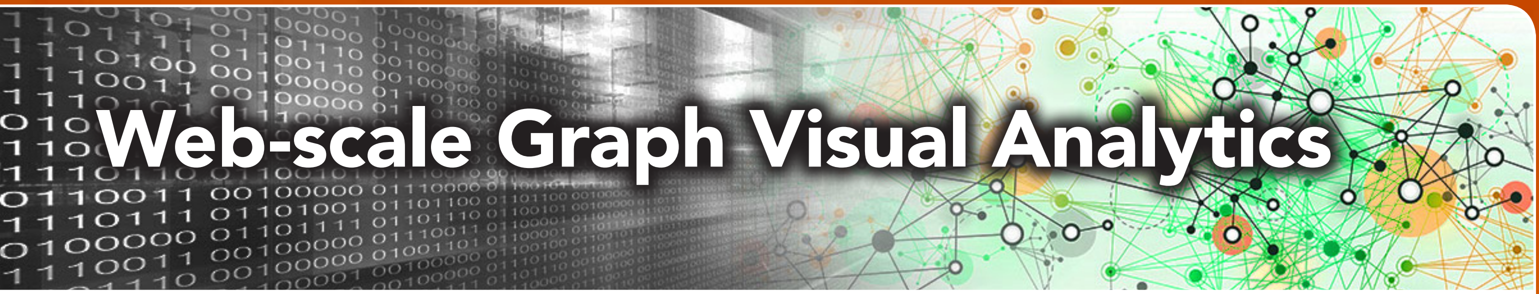
HIGH PERFORMANCE DATA ANALYTICS

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



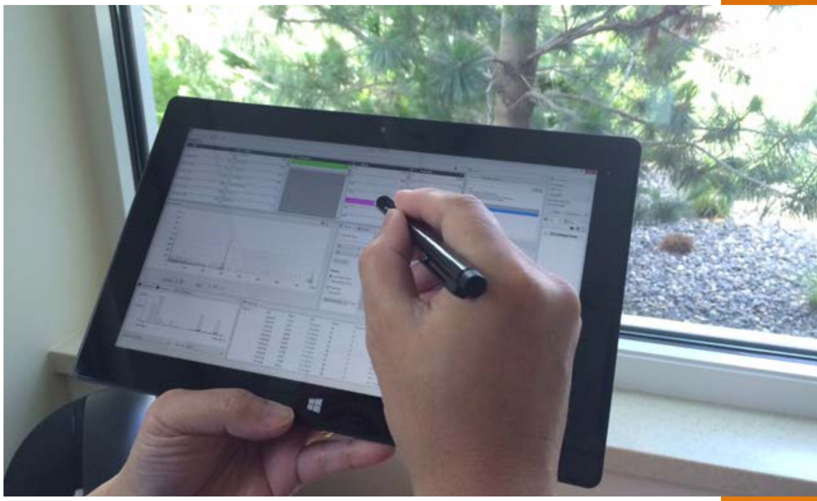
Web-scale Graph Visual Analytics

CHALLENGE

In today's big data era, a web-scale graph is often described as a graph with approximately one trillion edges and roughly 50 billion vertices. Exploring web-scale graphs is highly relevant in cloud-computing and cyber-analytics communities led by major internet vendors, including Google, Facebook, Microsoft, and Amazon. Although cutting-edge graph exploration technologies, such as Giraph and Pregel, have been reported in the literature, none use visualization in their analytical solutions.

Visually exploring a web-scale graph will require overcoming scalability challenges in size, cognition, visualization, and computation. This project's goal is to develop a visual paradigm and a client-server system that address these challenges using a combination of innovative visual analytics techniques in the front-end client and a high-performance database in the back-end server. When the technology is fully developed and integrated, the system is envisioned to allow users to query web-scale graphs in near interactive response time and visually explore the graph data on commodity level hardware.

> Exploring a web-scale graph that will overcome scalability challenges in size, cognition, visualization, and computation.



Running the big graph visual analytics system on a Microsoft Surface Pro tablet.

CURRENT PRACTICE

For nearly two decades, Shneiderman's visual information-seeking mantra, which states "overview first, zoom and filter, then details-on-demand," has influenced the thinking and development of the entire data visualization community. Its foundational design promotes a top-down approach of visualizing the whole before the parts, which is not practical for web-scale graphs. Our *peek-and-filter* graph visualization concept represents a paradigm shift from conventional method of using an overview visualization to guide data analysis to the new belief of first determining the right big data mix before conducting visual analytics.

U.S. DEPARTMENT OF
ENERGY

TECHNICAL APPROACH

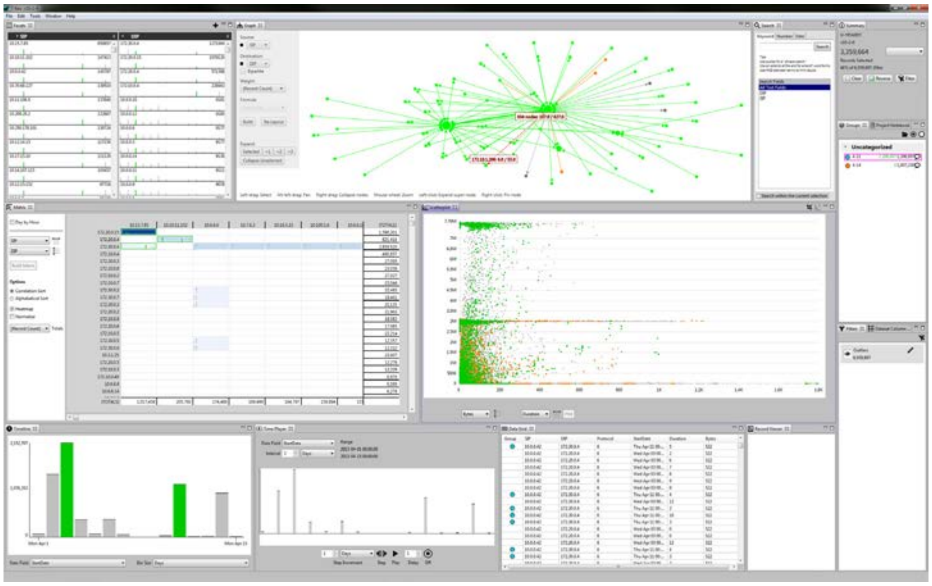
We are developing a web-scale graph visual analytics tool, known as T.Rex, to explore big graphs of unprecedented size. Working in concert with a high-performance database in the system back end, T.Rex provides a glimpse of the underlying data and allows users to decide if they want to commit to a potentially costly visualization step (refer to screenshot). A YouTube video that showcases the user interface and potential applications of T.Rex is available at www.youtube.com/watch?v=GSPkAGREO2E.

We have experimented with a number of database options, including Graph Engine for Multithreaded Systems (GEMS) and PostgreSQL, to power T.Rex's back end. GEMS is a property graph database supporting both traditional relational database system (Structured Query Language, or SQL) and graph-oriented operations such as edge traversal. Both have strengths and weaknesses for different graph exploration applications. A version of T.Rex that uses PostgreSQL as the database server was delivered in early 2016 and is available for use on federally funded projects through a "Government Use Agreement" license issued by Pacific Northwest National Laboratory.

A major research and development (R&D) effort is under way to include a multilevel hierarchical visualization technique focusing on the graph structures, such as cliques, instead of edges and vertices of a web-scale graph. The new addition is implemented in Scala using Apache Spark running on high-performance computing clusters both at PNNL and Mississippi State University (MSU). We are also planning to integrate GEMS with T.Rex, provided that all critical components are available within the project development timeframe. When the system is fully established, exploring big graphs using T.Rex and GEMS on a Linux system compatible to PUMA will scale to O (1T) graph edges.

IMPACT

The peek-and-filter paradigm will potentially influence the future R&D of not just the visualization of web-scale graphs, but also other types of big data sets that are too big to fit in one node of a computer cluster. A successful integration of T.Rex and GEMS will allow PNNL to address critical sponsor needs and continue to be an innovative leader in data visualization.



A screenshot of T.Rex exploring a data set with 6,939,698 NetFlow records or 627,573,995 graph edges.

Contacts

Pak Chung Wong
Principal Investigator
(509) 372-4764
Pak.Wong@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Song Zhang
(662) 325-7510
szhang@cse.msstate.edu

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

MISSISSIPPI STATE
UNIVERSITY™