Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

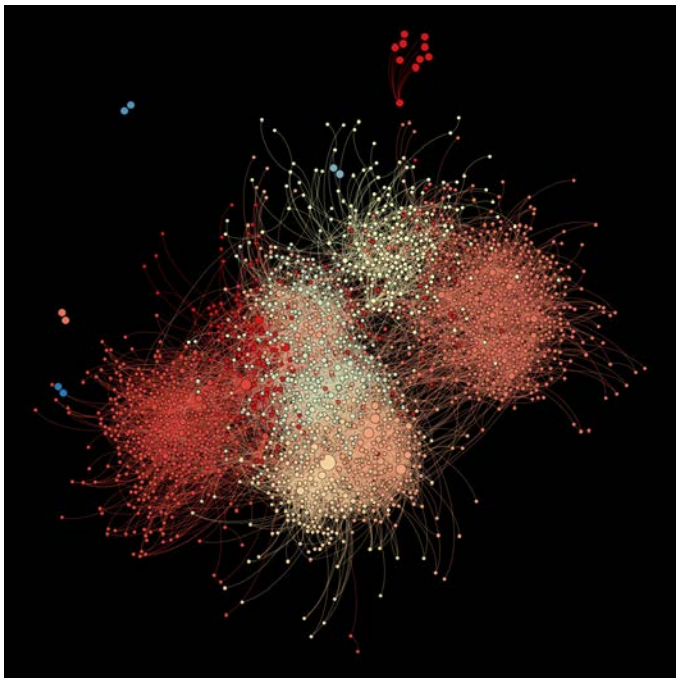
Performance Modeling on Property Graphs

CHALLENGE

The term "property graph" has come to represent directed, multi-relational graphs with collections of arbitrary key-value pairs associated with nodes and edges. A property graph can be used to model data sets, such as network traffic flows where nodes represent Internet protocol (IP) addresses and edges indicate different types of interactions between the hosts. In addition, every edge in the graph may contain key-value pairs to capture the amount of exchanged bytes and session durations. This research will address two major challenges with extreme-scale property graphs:

- Challenge A. Developing graph generators
- Challenge B. Quantitatively modeling the performance of key algorithms, such as graph search.

Developing new modeling capabilities for property graphs that include graph generation algorithms and quantitative understanding of graph queries at extreme scales.



A simulated property graph using the Multiplicative Attribute Graph Generation model.

CURRENT PRACTICE

Challenge A. Over the past 10 years, the Semantic Web and database community has developed a large collection of synthetic data generators and resource description framework (RDF) benchmarks. Conversely, there are well-known graph generators that generate random graphs with statistical models, such as power-law distributions and Kronecker models. However, algorithms to generate property graphs by combining both complex network models (e.g., for realistic simulating of host-host interactions) and approaches for generating relational/Semantic Web data with a

specified schema and data distribution (e.g., source and destination ports, number of bytes, etc.) remain a nascent area.

Challenge B. Traditionally, most graph search techniques have relied on the ability to index the graph by labels or node-level patterns. However, many of these indexing techniques become infeasible at the extreme scale because of their high computational complexity. Therefore, greater emphasis is placed on using algorithmic approaches, such as query decomposition, the intelligent partitioning to develop scalable runtime systems, to address irregular computation issues. Another promising area is approximate query processing, where the goal is to return the top-ranking answers within a specified time limit.

TECHNICAL APPROACH

Understanding the interplay between graph structure and the distribution of node or edge labels and attributes is central to solving the seemingly disjointed problems already described.

We seek a two-phased approach to generating property graphs. The first phase will use traditional Kronecker-like models to simulate the underlying graph structure. Next, we will model the nodes as multivariate random variables, where each random variable represents a combination of degree, node label, and other attributes. Once the graph structure is simulated, we will iterate through every node or edge and use the underlying generative model to assign an appropriate combination of the labels and attributes.

Next, we plan to study the performance of graph queries through the controlled variation of multiple factors: variability in the property graph structure, distributed graph partitioning strategies, and variation in queries via permuting across multiple categories in terms of structure and selectivity. We plan to build a predictive model for query performance through extraction of features of the query graphs and the target database. Identification of efficiently computable features that yields a model with an acceptable level of accuracy will be a major focus of this research.

IMPACT

Property graphs are becoming increasingly popular as more applications produce data that contain heterogeneous interaction between entities, and include significant node or edge label attribute information that cannot be efficiently represented by reification-like approaches. The ability to generate massive property graphs based on a small sample data set will be a critical tool for any research that requires testing or benchmarking on property graphs. Our graph query performance model will provide an important optimization capability for any graph database. Given a data graph and a set of target queries, the model can guide strategies for a faster search or enable approximate query processing by determining the most relevant subgraphs to search.

Contacts

Sutanay Choudhury
Principal Investigator
(509) 375-3978
sutanay.choudhury@pnnl.gov

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

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965