Benchmarking Heterogeneous Graph Data

Amarnath Gupta
Univ. of California San Diego

What We Do

- Semantic Information Integration in Scientific Domains
  - The Neuroscience Information Framework (www.neuinfo.org)
    - Several hundred data sources (relational, XML, RDF, OWL, domain-specific, ….) and counting
    - Several inter-mapped ontologies
    - Data mapped to ontology terms and relationships when possible
  - Can be viewed as an edge-labeled graph
    - ~350M nodes, ~3B edges, 2000 edge labels
  - Expecting a 50-75 fold growth over the next couple of years

Multiple User Groups

What We Observe

- The graph is far from homogeneous
- Can be decomposed into connected modules that have different structural and operational characteristics

Component	Operations
Ontology	Transitive closure, Center-piece Subgraph, Clustering
Data Graph	Subgraph extraction, Path queries, Over-representation analysis
Connectivity	Centrality properties, Network features, Motif discovery, network similarity analysis, Path queries
Pathways	Traversal, Subgraph extraction, Graph pattern queries
Citation Graph	Hub and influence analysis, Data connectivity analysis
Text Graph	Centrality properties, frequent subgraph mining

What We Want

- A data generator that will produce a heterogeneous graph of different sizes and types, component complexities and inter-component connectivity patterns
- A set of benchmark queries (and variants) that will exercise a variety of common operations on generated data sets.
  - Within components
  - Across components
- A query workload
- Bulk updates

Fake Example Query

```
SELECT neighborhood(N, 3) AS B
FROM G.C2
WHERE N = induced-subgraph(
        SELECT node from G.C2
        WHERE node.type = X
        AND node.label != Y)
AND B.edge.label IN (L1,L2,L3)
```

A Starting Point

- Consider a spec like this:
  - **WITH** Table1,Table2,Table3 as ***NodeLabels***,
    Table4, Table5, Table6 as ***EdgeLabels***

    GENERATE Graph G (Component C1, C2, C3)

    WHERE C1.NodeLabel in Table1, C2.NodeLabel …

    AND C1.EdgeLabel in Table4, …

    AND G.AvgNodeCount = 500000000

    AND C1.type = rootedDAG (avgFanOut=5, avgFanIn=3, numLevels=20, …)

    AND C1.edgeFraction = 0.2

    …

    AND G.AvgConnectionCount(C1,C2) = 400000

    …

- Graph Types
  - standard
  - unLabeledPowerLaw
  - multiHubGraph
- Variables to Specify
  - Min and max degrees
  - Coefficient of Power Law
  - Number of hub nodes
  - Distribution of labels
  - Connectivity across components

- Graph 500
  - SCALE N:# nodes M = 2^N
  - EDGEFACTOR E: # of edges = E*M

A Proposition

- No standards and no data generating mechanisms
- We can provide sample data to help simulate the multi-component network [1]
- Can we form a working group to specify the benchmark DB and a DB-generating mechanism by the next workshop?

- Reference

1. S. Duan, A. Kementsietsidis, K. Srinivas, O. Udrea: "Apples and oranges: a comparison of RDF benchmarks and real RDF datasets". ACM SIGMOD Conference 2011: 145-156.