Graph Storage and Indexing

• deduplicated storage
• compact graph of graphs

Deduplicated Graph Distribution

Team Members

Sean R. Spillane (seans@cs.albany.edu) and Jeong-Hyon Hwang (jhh@cs.albany.edu)
Jeremy Bimbaum
Daniel Bokser
Daniel Kemp
Alan G. Labouseur
Paul W. Olsen, Jr.
Jayadevan Vijayan

Key Features
To efficiently manage data that represent large, evolving networks, G* offers:
• distributed, deduplicated storage of graphs
• compact graph indexing
• sophisticated graph queries
• shared computation across graphs

Applications
Discovering the characteristics of evolving networks is essential for:
• marketing
• sociology
• national security
• transportation
• fraud detection in financial markets
• epidemiology
• pharmacology
... many other areas.

Data Sets
The benefits of G* can be demonstrated using real-world and synthetic data sets, such as:
• Twitter messages
• Yahoo! server logs
• DBLP citation and coauthorship data
• synthetic binary trees

Queries
G* supports a variety of queries that find:
• the variation of the average degree over graphs (as shown above)
• the degree distribution for each graph
• the clustering coefficient distribution for each graph
• the distribution of geodesic distances from a vertex to all other vertices
• the variation of a vertex’s centrality
• vertices with the largest increase in centrality
• the size of the largest connected component for each graph

Graph Query Execution

G* vs. PostgreSQL
- 11 cumulative Twitter graphs (each with 1K new edges)
- clustering coefficients
- vertex degree
- vertex centrality
- query time (sec)

G* vs. Giraph
- 10 cumulative binary trees (each with 1M new edges)
- geodesic distances from root to all vertices

Experimental Results

This work is supported by NSF CAREER award IIS-1149372

The Graph Database System

G*