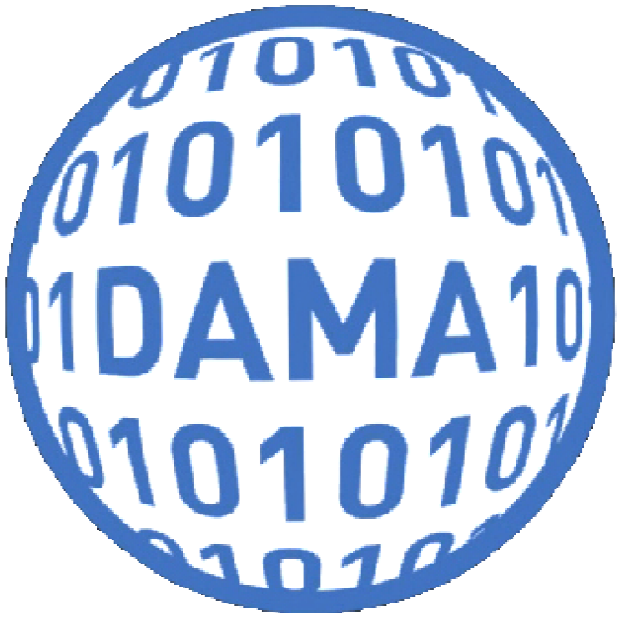
DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA



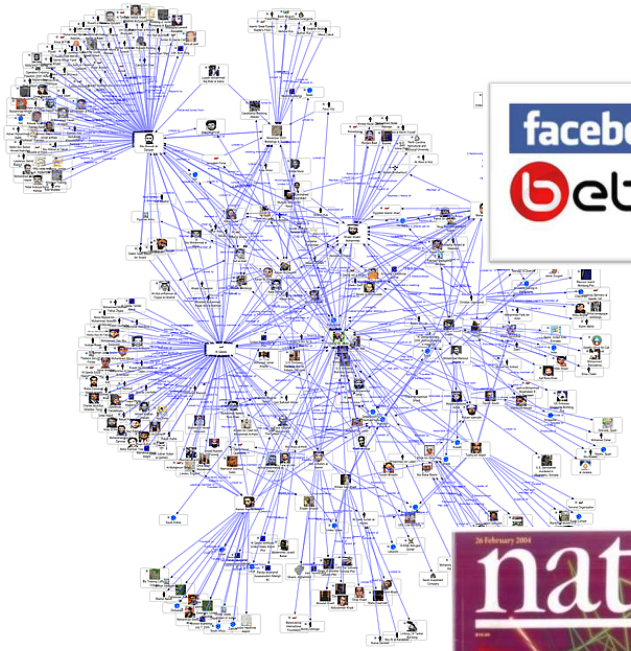
A Discussion on the Design of Graph Database Benchmarks

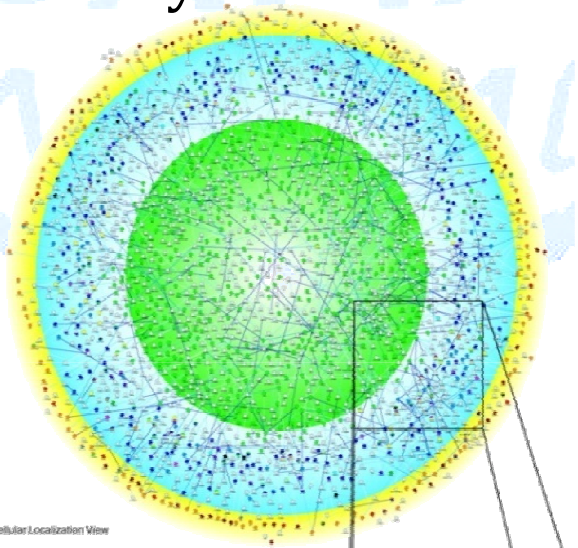
David Dominguez-Sal, Norbert Martínez, Victor Muntes, Pere Baleta, Josep Lluis Larriba

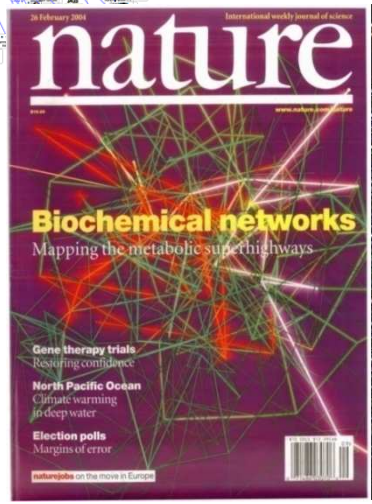
TPCTC 2010
Singapore

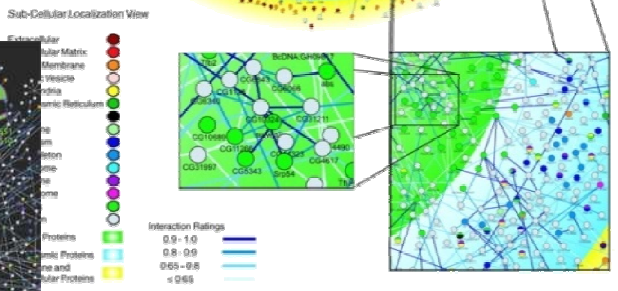
Motivation

- Growing volumes of graph data to analyze









Motivation

- Emerging market
  - Many new graph libraries
  - Neo4j, HypergraphDB, Pregel, Jena-RDF, DEX, etc.
- Performance?
  - Benchmark graph databases
- Other benchmarks not suitable
  - Relational, object oriented, XML, etc.
- Few proposals available
  - HPC-SGAB (Bader et al.)

Objectives

- Survey of graph applications with large data volumes
- Classify graph applications
  - Datasets
  - Operations
- Set GDB benchmarking as an open discussion topic

1. Introduction
2. Graph description
3. Graph operations
4. Experimental setting

Representative areas

1. Social graphs
   - Relations generated explicetely by human interactions.
   - E.g. Facebook, flickr, citation author networks...

2. Biological graphs
   - Relation defined by observations on nature
   - E.g. Protein to protein interaction, food web chain, biochemical reaction

Representative areas

3. Routing
   - Relations are physical (usually 2D)
   - E.g. Road routing, communication networks, real time traffic analysis.

4. Recommendation
   - Mixed information sources to mine
   - Eg: product recommendation, advertising…

Graph description

- Attributes
  - Nodes, edges (e.g. weight).
  - Identifiers
- Directed / Undirected
- Labeling (Typing)
- Multigraphs
- Hypergraphs
  - Hyperedges may be modeled as special nodes

1. Introduction
2. Graph description
3. Graph operations
4. Experimental setting

Graph operations

- Basic analysis:
  - Get node/edge
  - Get attributes from a node or an edge
  - Get neighbors
  - Node degree
- Basic transformations
  - Add/delete node/edge
  - Add/delete/update attribute

Graph operations

- High level operations
  - Traversals
  - Component analysis
  - Communities
  - Graph analysis (statistics)
  - Centrality measures
  - Pattern matching
  - Anonymization

Operation categorization

- Transformation / Analysis
- Cascaded access
  - At least depth 2 (friends of my friends)
- Scale
  - Global, neighborhood
- Attributes
  - Nodes, edges, none
- Result
  - Graph, aggregated results, sets.

Summary of graph operations

Group	Operation	Social Network	Protein Interaction	Recom-mendation	Routing	Analytical	Cascaded	Scale	Attr.	Result
Generic operations										
General Atomic / Local Information Extraction	Get node/edge	+	+	+	+	Yes	No	Neigh.	No	Set
	Get attribute of node/edge	+	+	+	+	Yes	No	Neigh.	No	Set
	Get neighborhood	+	+	+	+	Yes	No	Neigh.	No	Set
	Node degree	+	+	+	+	Yes	No	Neigh.	No	Agr.
General Atomic Transformations	Add/Delete node/edge	+	+	+	+	No	No	Neigh.	No	Set
	Add/Delete/Update attrib.	+	+	+	+	No	No	Neigh.	E/N	Set
Application dependent operations										
Traversals	(Constrained) Shortest Path	+	+		+	Yes	Yes	Glob.	Edge	Graph
	k-hops	+		+	+	Yes	Yes	G/N	No	Graph
Graph Analysis	Hop-Plot	+				Yes	No	Glob.	No	Agr.
	Diameter	+	+			Yes	Yes	Glob.	Edge	Set
	Eccentricity	+				Yes	Yes	Glob.	Edge	Agr.
	Density	+	+			Yes	No	Glob.	No	Agr.
	Clustering coefficient	+				Yes	Yes	Glob.	No	Agr.
Components	Connected Components	+	+			Yes	Yes	Glob.	No	Graph
	Bridges	+	+		+	Yes	Yes	Glob.	No	Set
	Cohesion	+				Yes	Yes	Glob.	No	Set
Communities	Dendrogram	+				Yes	Yes	Glob.	No	Graph
	Max-flow min-cut	+				Yes	Yes	Glob.	Edge	Graph
	Clustering	+	+	+		Yes	Yes	Glob.	No	Graph
Centrality Measures	Degree Centrality	+		+		Yes	No	Glob.	No	Set
	Closeness Centrality	+		+		Yes	Yes	Glob.	No	Set
	Betweenness Centrality	+		+		Yes	Yes	Glob.	No	Set
Pattern Matching	Graph/Subgraph Matching	+	+			Yes	Yes	Neigh.	No	Graph
Graph Anonymization	k-degree Anonym.	+		+		Yes	No	Glob.	No	Graph
	k-neighborhood Anonym.	+		+		Yes	Yes	Glob.	No	Graph
Other Operations (Similarity, ranking,...)	Structural Equivalence	+				Yes	Yes	Glob.	No	Graph
	PageRank			+		Yes	No	Glob.	Node	Set

Table 1. Graph Operations, Areas of Interest and Categorization

1. Introduction
2. Graph description
3. Graph operations
4. Experimental setting

Experimental setting

- Configuration and setup
  - Data partitioning, indexing, redundancy, data reorganization,
  - ACID? Isolation? Eventual consistency?
- Experimental process
  - Warmp up, query sequence, sampling procedure
- Measures
  - Simple but adapted to the audience
  - Eg: Load time, response time, throghput, image size, power, price/throughput, etc.
  - Adapted to graph TEPS, query completeness vs time

Conclusions

- Graph databases is an emerging market
  - Large volumes of graph data available to analyze.
- Many applications appearing
  - Benchmark comparison
- Graphs are varied and its applications differ, but they have many shared aspects

Conclusions

- Expectations of a generic graph benchmark:
  - Attributed, labeled (types), directed, multigraph.
  - Significant set of cascaded and graph result operations
  - Definition of experimental process
- Candidate scenario: Social networks
  - Large datasets, variety of operations, industrial interest.
- Future work
  - Materialize the benchmark

Thanks!

Questions