Social Network Benchmark Task Force

3rd TUC Meeting
London, 19 November 2013

Social Network Benchmark (SNB)

- Designed for evaluating a broad range of technologies for tackling RDF and graph database workloads
- Scenario:
  - Understandable to a large audience
  - Cover the complete range of interesting and realistic challenges
  - Same input (graph data) for different challenges (query workloads)

Why Social Network Analysis?

- Intuitive: everybody knows what a SN is
- SNs can be easily represented as a graph
- Different scales: from small to very large SNs
- Multiple query needs:
  - interactive, analytical, transactional
- Multiple types of uses:
  - marketing, recommendation, social interactions, fraud detection, ...

SNB - Audience

- For end users facing graph processing tasks
  - recognizable scenario to compare merits of different products and technologies
- For vendors of graph database technology
  - checklist of features and performance characteristics
- For researchers, both industrial and academic
  - challenges in multiple choke-point areas such as query optimization, (distributed) graph analysis, transactional throughput

SNB - Workloads

- Three distinct benchmarks:
  - **On-Line**
    - tests a system's throughput with relatively simple queries with concurrent updates
  - **Business Intelligence**
    - consists of complex structured queries for analyzing online behavior
  - **Graph Analytics**
    - tests the functionality and scalability on most of the data as a single operation

SNB - Systems

- Graph database systems:
  - e.g. Neo4j, InfiniteGraph, DEX, Titan
- Graph programming frameworks:
  - e.g. Giraph, Signal/Collect, Graphlab, Green Marl
- RDF database systems:
  - e.g. OWLIM, Virtuoso, BigData, Jena TDB, Stardog, Allegrograph
- Relational database systems
  - e.g. Postgres, MySQL, Oracle, DB2, SQLserver, Virtuoso, MonetDB, Vectorwise, Vertica

SNB – Expected Results

- Four main elements:
  - *data schema*: defines the structure of the data
  - *workload*: defines the set of operations to perform
  - *performance metrics*: used to measure (quantitatively) the performance of the systems
  - *execution rules*: defined to assure that the results from different executions of the benchmark are valid and comparable
- Software is open source (GitHub)
  - data generator, query drivers, validation tools, ...

SNB Task Force

- University
  - VUA - The Vrije Universiteit Amsterdam
  - UPC - Universitat Politècnica de Catalunya
  - TUM - Technische Universtität München
- Industry
  - RDF
    - OpenLink Software (Virtuoso)
  - Graph Databases
    - Neo Technology (Neo4J)
    - Sparsity Technology (DEX)

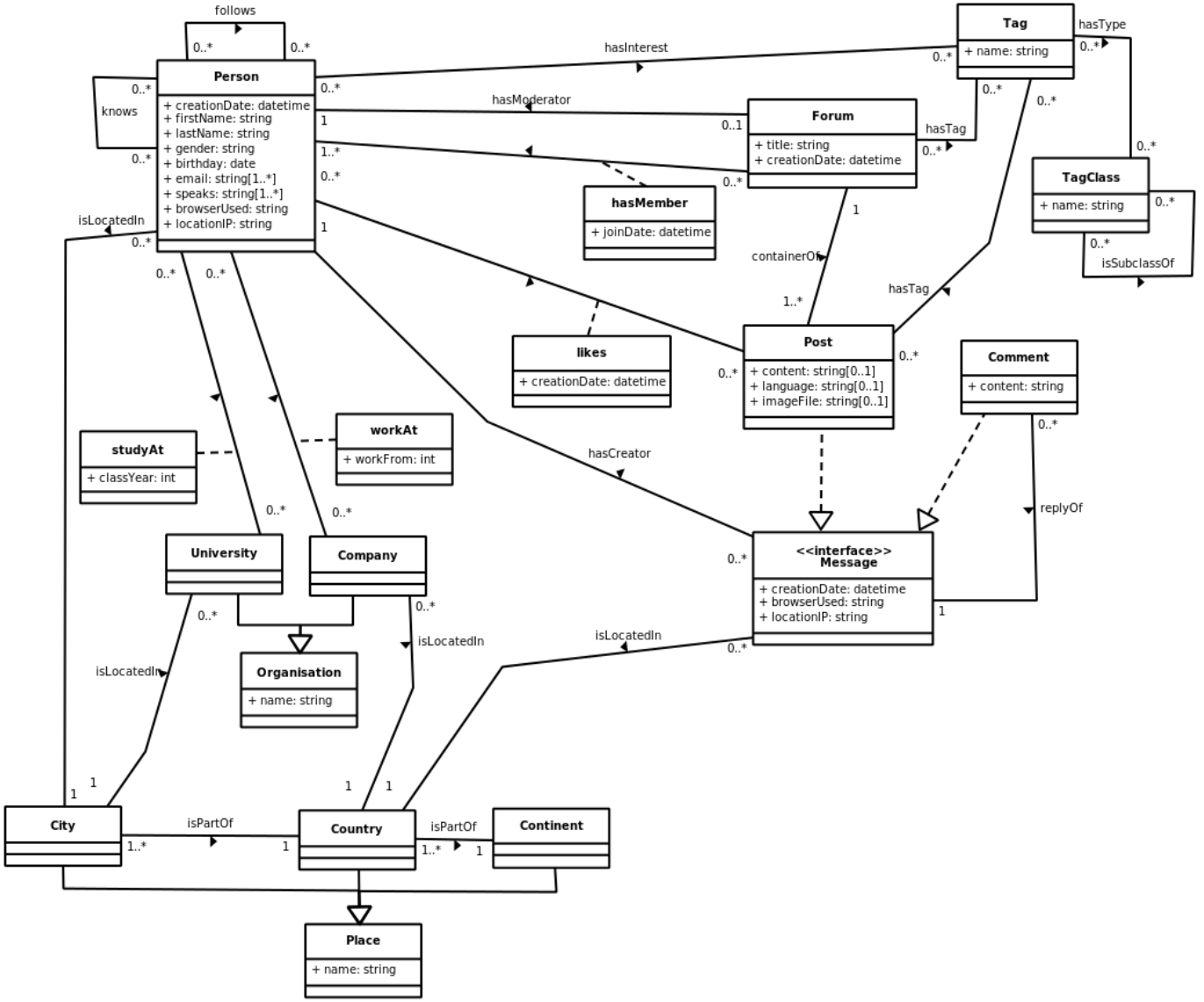
SNB Activities

- 2 TUC meetings: Barcelona and Munich
- Dataset Generator and Interactive Query Set
- 4 scientific papers covering technical aspects of benchmark development
- Organization of the First Intl. Workshop on Graph Data Management Experiences and Systems (GRADES), co-located with SIGMOD/PODS 2013
- Presentation at GraphLab 2013, an event focusing on graph programming frameworks

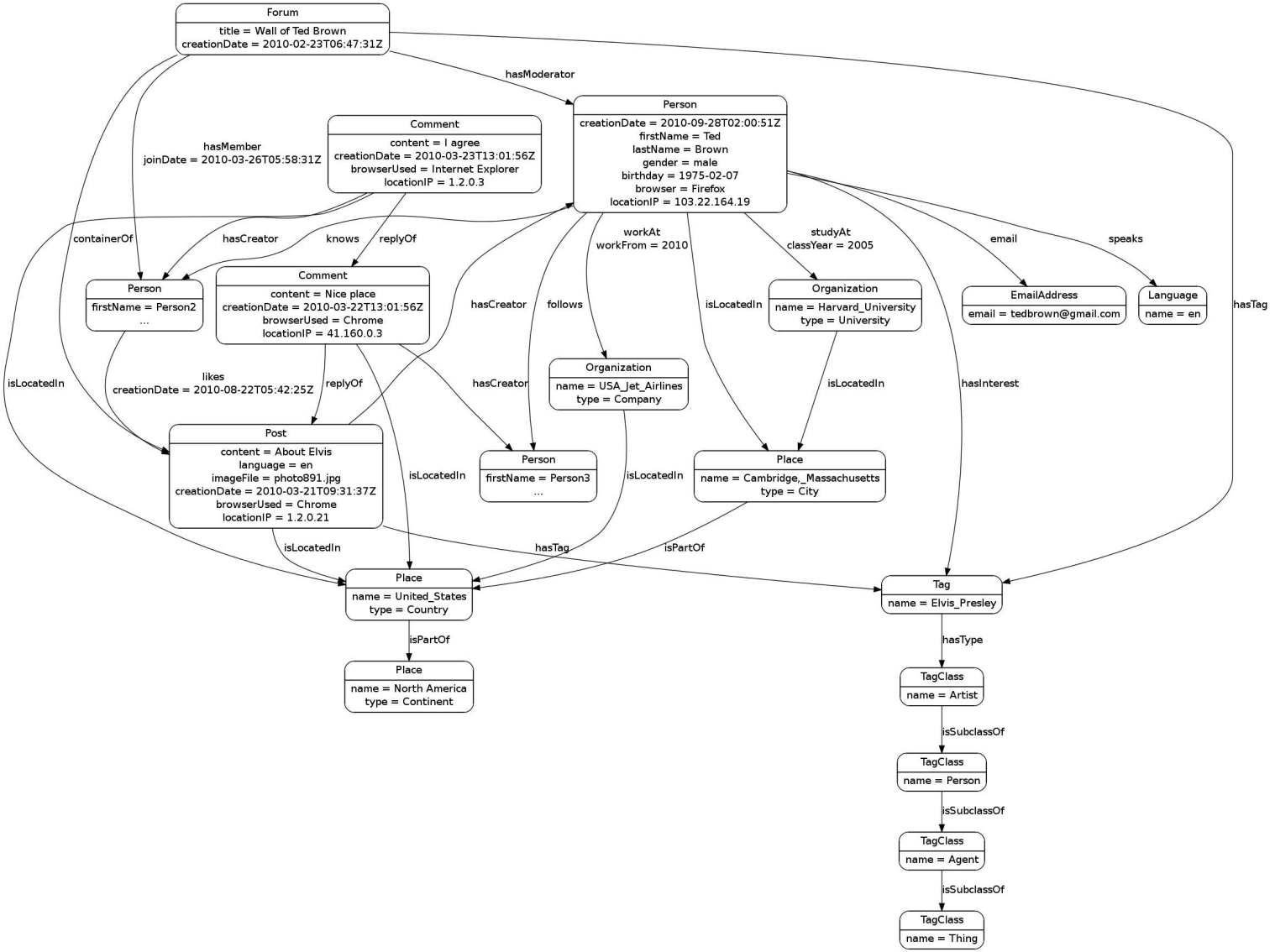
SNDG - SNB Data Generator

- Correlated Property Graph
- Mimics the characteristics of real SN data
- Based on SIB–S3G2 Social Graph Generator
  - *property dictionaries* extracted from DBPedia with specific ranking and probability density functions
  - *subgraph generation*: new nodes and new edges in one single pass (based on degree distributions)
- MapReduce
- Outputs RDF and CSV

SNDG – Data Schema



SNDG Graph Example



LDBC
Linked Data Benchmark Council

SNDG Statistics (100K/1Y)

Group	Statistic	Value
Settings	Number of users (*Person* instances)	100,000
	Number of years	1
Global	Nodes	80,767,146
	Edges	350,352,746
	Attribute Values	500,108,979
Metrics	Largest connected component (community)	99.78%
	Average path length (small world)	3.93
	Average clustering coefficient (transitivity)	0.11
	Largest distance between two nodes (diameter)	11
Knows relationship	Edges	2,887,796
	Diameter	6

Choke Point Analysis

- Carried out using the expertise of database architects and researchers
- Goal: identify the most important technical challenges that should be tested by the queries:
  - scale of data
  - different platform types, including commodity server clusters and shared memory scale-up solutions
  - stress parallelism and locality
  - ...

Choke Points

Group	Description	#
CP1	Aggregation Performance. Performance of aggregate calculations.	7
CP2	Join Performance. Voluminous joins, with or without selections.	7
CP3	Data Access Locality. Non-full-scan access to (correlated) table data.	5
CP4	Expression Calculation. Efficiency in evaluating (complex) expressions.	11
CP5	Correlated Subqueries. Efficiently handling dependent subqueries.	3
CP6	Parallelism and Concurrency. Making use of parallel computing resources.	3
CP7	RDF and Graph Specifics	3
CP8	Update Concurrency	3
CP9	I/O	1

SNB Interactive Query Set

- Tests system throughput with relatively simple queries and concurrent updates
- First version: twelve read-only queries
  - 16 choke points
- Example: Q11

  Find a friend of the specified person, or a friend of his friend (excluding the specified person), who has long worked in a company in a specified country. Sort ascending by start date, and then ascending by person

SNB Query Drivers

- QGEN - BIBM
  - based on BSBM
  - capability of running multiple clients concurrently
  - comparison of the result sets
  - warm-up runs
- LDBC_DRIVER
  - started out as a fork of YCSB
  - key-value model to represent a graph model

Interactive Query Set Experiments

- Virtuoso (RDF)
  - 100k users during 3 years period (3.3 billion triples)
  - Ten query mixes
  - 4 x Intel Xeon 2.30GHz CPU, 193 GB of RAM
- DEX (Graph Database)
  - Validation setup: 10k users during 3 years (19GB)
  - Validation query set and parameters
  - 2 x Intel Xeon 2.40Ghz CPU, 128 GB of RAM

Preliminary Results

- Some queries could not be considered as truly interactive (e.g. Q4, Q5 and Q9)
  - still all queries are very interesting challenges
- ”Irregular” data distribution reflecting the reality of the SN
  - but complicates the selection of query parameters
- Both systems have identified some of their “internal” implementation choke points
  - some optimizations implemented and tested

Future Work (2nd year)

- Workloads
  - interactive updates (transactional)
  - new BI and Graph Analytical
  - substitution parameters
- Data Generator
  - improve dictionaries and distributions for BI
- Query Drivers
- Scale factors and dataset (SN graph) validation
- Auditing rules