



Social Network Analytics Task Force

2nd TUC Meeting
Munich, 22-23 April 2013

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

SNA: Types of Uses

- Marketing
- Community management
- Recommendation
- Social Interactions
- Security & Fraud Detection
- Business Intelligence: churn analytics, CRM...

Why Graph Databases?

- A SN is usually represented as a network (a graph)
 - Entities are the nodes (Person, Post, ...)
 - Relationships are the edges (Friend, Likes, ...)
- Graph databases:
 - information management systems that are able to perform graph-oriented operations efficiently

Social Intelligence Benchmark

- The SIB is designed for evaluating a broad range of technologies for tackling graph database workloads
- Scenario:
 - It should be understandable to a large audience
 - It should cover the complete range of interesting challenges
 - The query challenges in it should be realistic

SIB - 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

SIB - Workloads

- Three distinct benchmarks (workloads):
 - **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
- Each workload produces:
 - a single metric for performance at the given scale
 - a price/performance metric at the scale

SIB - 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

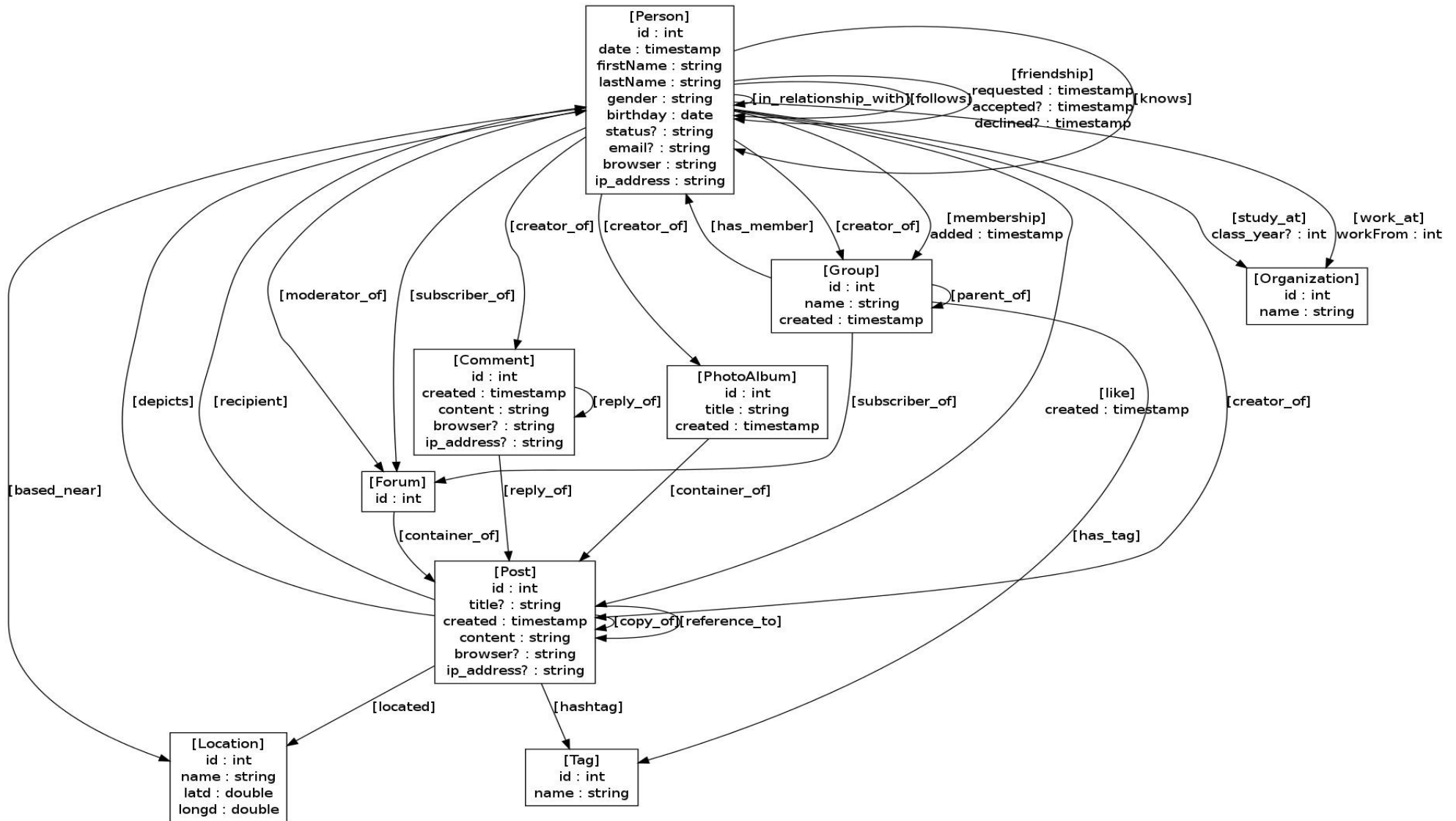
SIB - Design Objectives

- Rich coverage
- Modularity
- Reasonable implementation cost
- Relevant selection of challenges
- Reproducibility and documentation of results

SIB – Dataset Generator

- Based on SIB–S3G2 Social Graph Generator
- Synthetic generated data linked with the RDF datasets from the DBpedia knowledge base
- Entities:
 - Person, Group, Organization
 - Forum, Post, Comment, PhotoAlbum
 - Location, Tag

SIB – Graph Model (proposal)



SIB – Query Examples

- Interactive
 - Find top-10 suggested friends for a user
 - Show all photos posted by my friends that I was tagged in
 - Updates (transactional)
- Analysis
 - Post frequency distribution
 - Who got the most replies during 1st month of participation?
 - Person posting about X and Y with X and Y likes scores
- Algorithms
 - PageRank
 - Reachability / shortest path

Work in Progress

- Methodology for a system-independent specification of a Graph Logical Schema
- SN Graph Logical Schema
- SIB Data Generator
- Query Choke Points analysis
- Analysis of alternative benchmarks
 - LinkBench
- Analysis of benchmark platforms
 - YCSB

Team

- University
 - VUA - The Vrije Universiteit Amsterdam
 - UPC - Universitat Politècnica de Catalunya
 - TUM - Technische Universität München
- Industry
 - OGL - OpenLink Software
 - NEO - Neo Technology

User Community

- 1st TUC (Nov 2012)
 - ACCESO
 - MPG
 - ERA7 Bioinformatics
 - InnoQuant
- 2nd TUC (Apr 2013)
 - ...
 - ...
 - are you joining us?