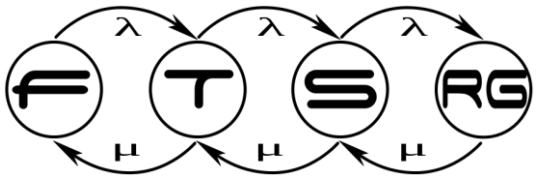
Budapest Neo4j Meetup – 2019/06/25

What makes graph queries difficult?

Gábor Szárnyas
szarnyas@mit.bme.hu

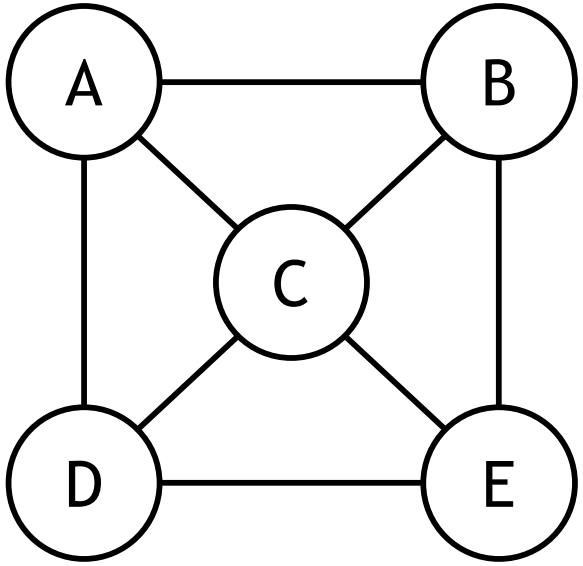


Hungarian Academy of Sciences

With contributions from Petra Várhegyi and Bálint Hegyi

The property graph data model

SIMPLE GRAPH



5 people

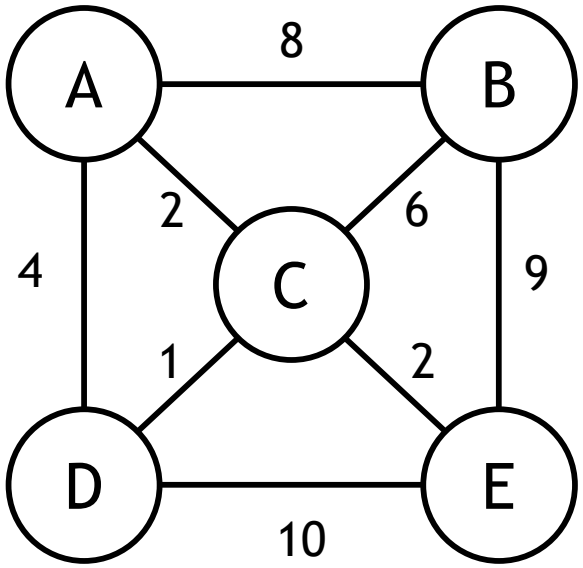
- Many of them know each other

This is a **simple graph**.

Algorithms:

- breadth-first search
- depth-first search
- PageRank
- connected components

ADD EDGE WEIGHTS



5 people

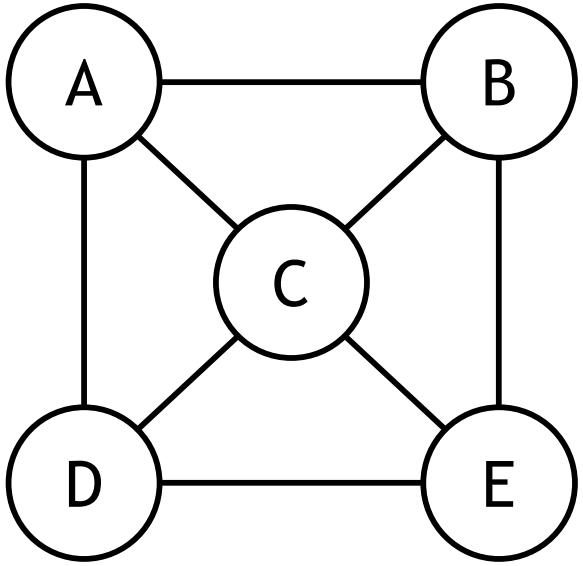
- Weight: communication cost

This is a **weighted graph.**

Algorithms:

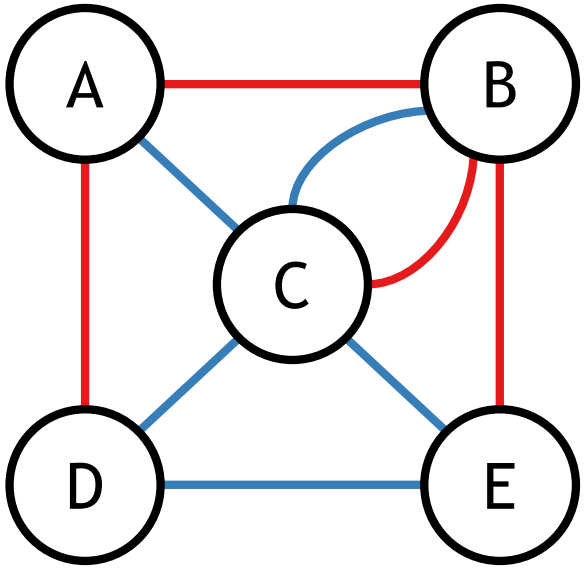
- shortest path algorithms
- max-flow

ADD EDGE TYPES



5 people

ADD EDGE TYPES



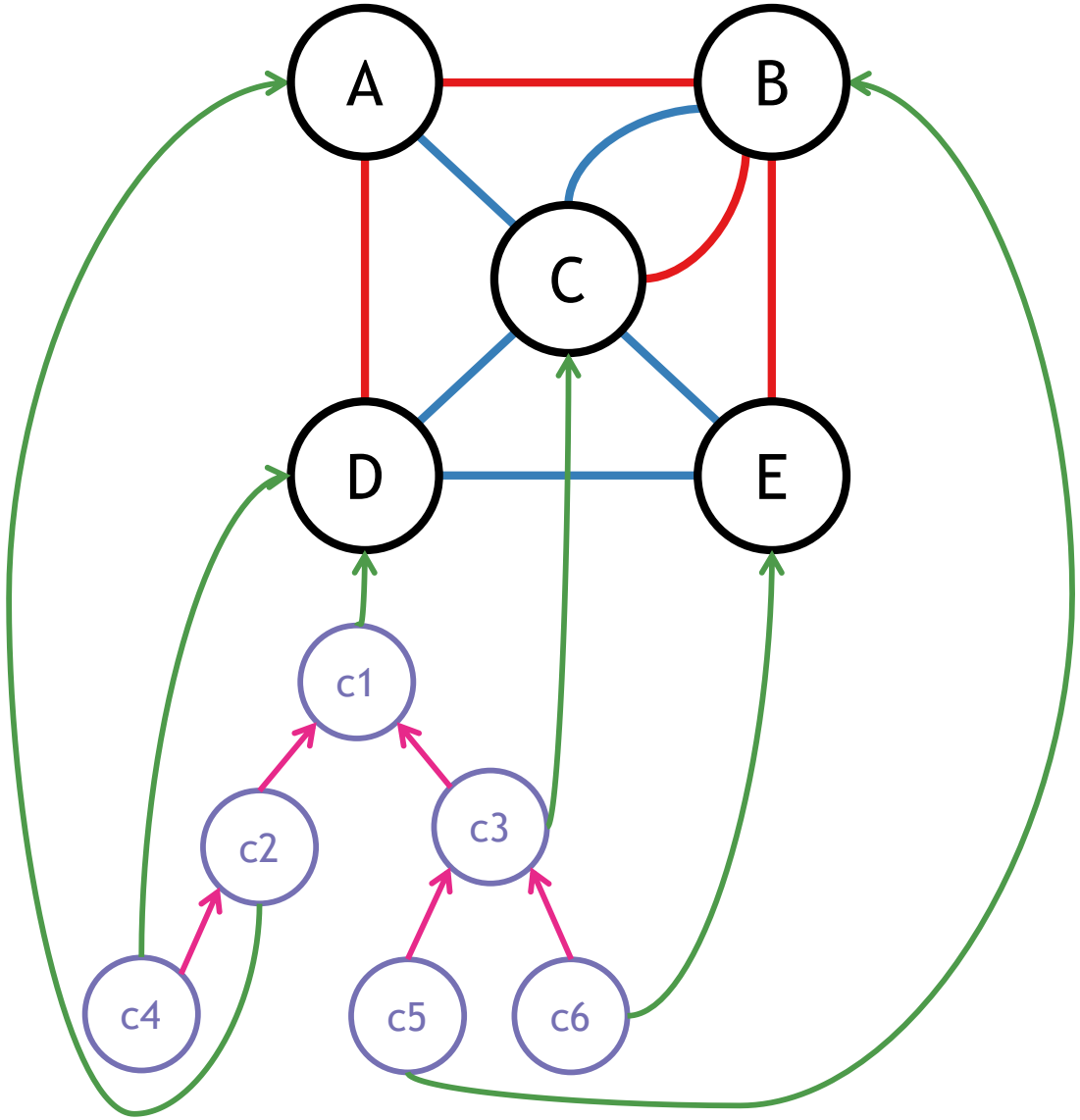
5 people

- Business partners
- Friends

Multiple edge types
but only a single node type.

This is an **edge-typed graph**.

ADD NODE AND EDGE TYPES



5 people ●

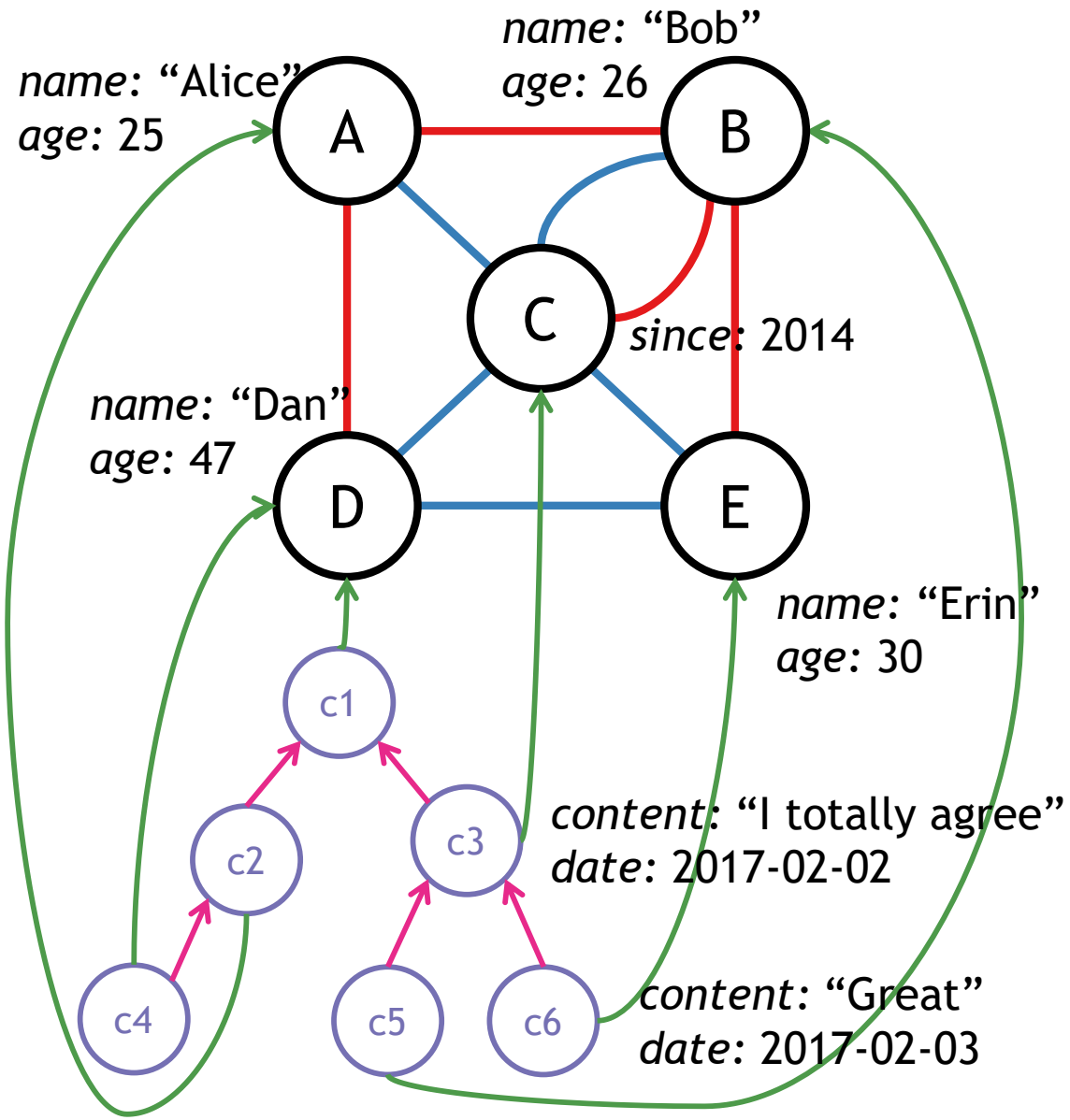
- Business partners
- Friends

6 comments ●

- Replying to another comment
- Authored by a given person

This is a **typed graph.**

ADD PROPERTIES



5 people ● - *name*, *age*

- Business partners
- Friends - *since*

6 comments ● - *content*, *date*

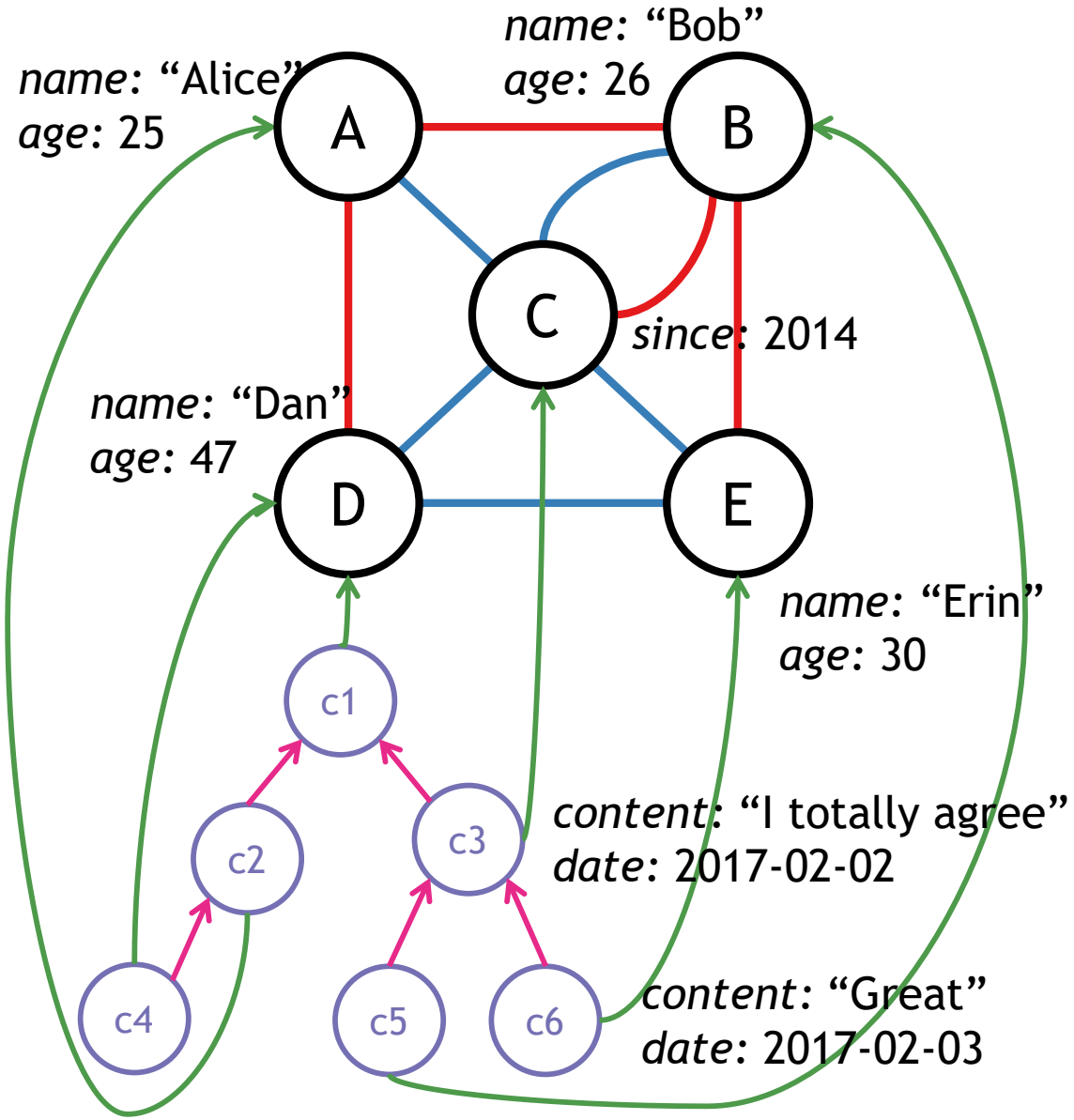
- Replying to another comment
- Authored by a given person

This is a **property graph.**

Similar to object-oriented data.

Graph processing: Queries and analytics

GRAPH QUERIES: LOCAL



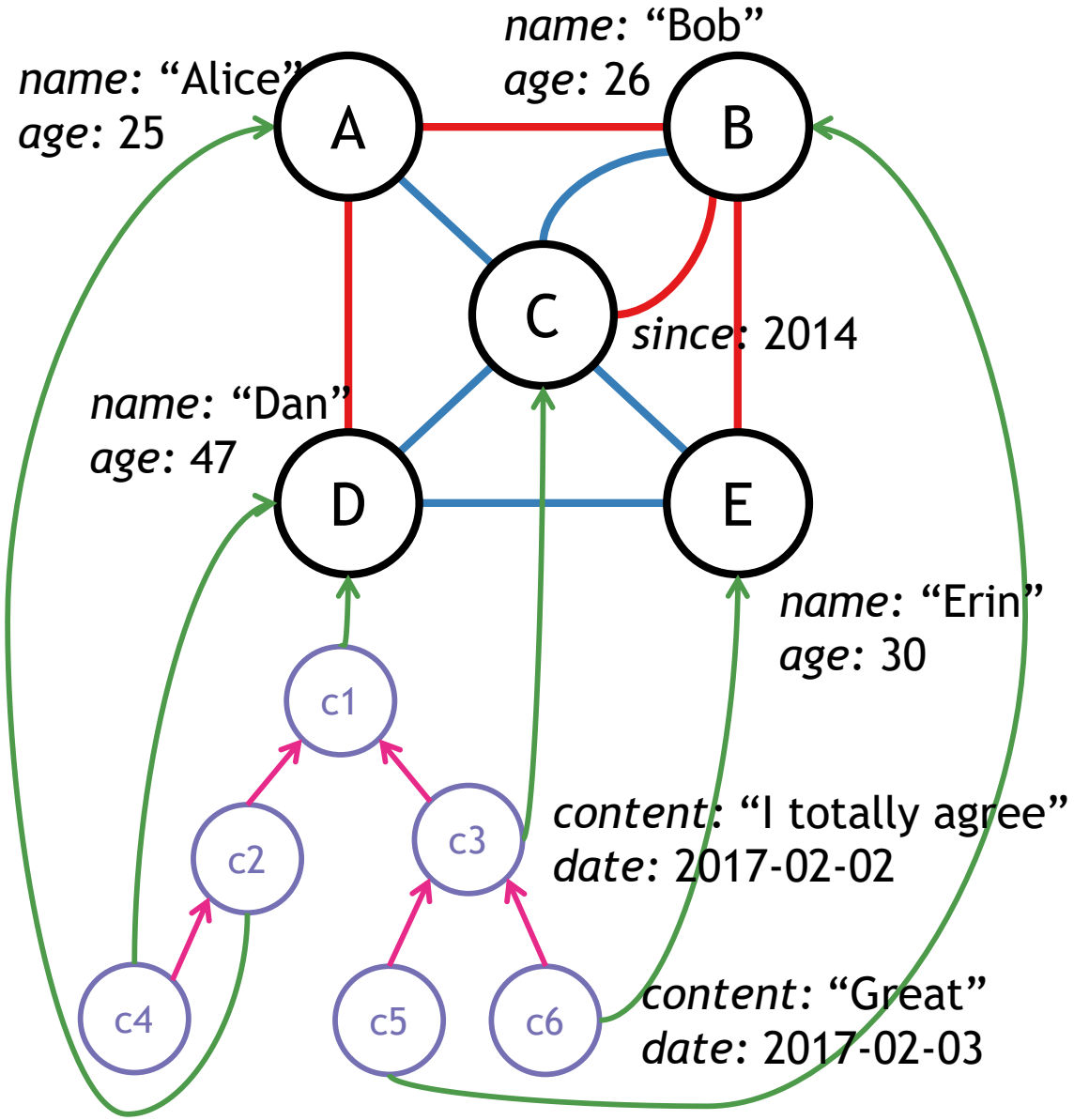
Local graph query:

Return "Dan" and his comments.

Well-researched topic.

Typical execution times are low.

GRAPH QUERIES: GLOBAL



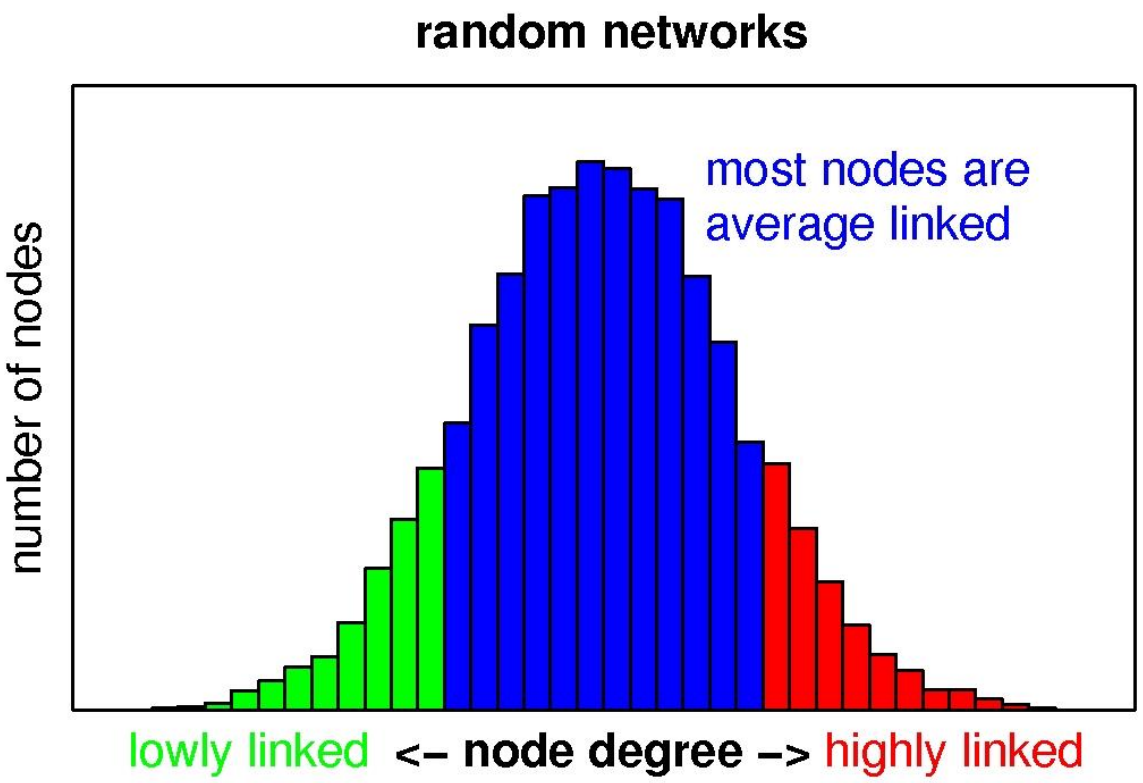
Global graph query:

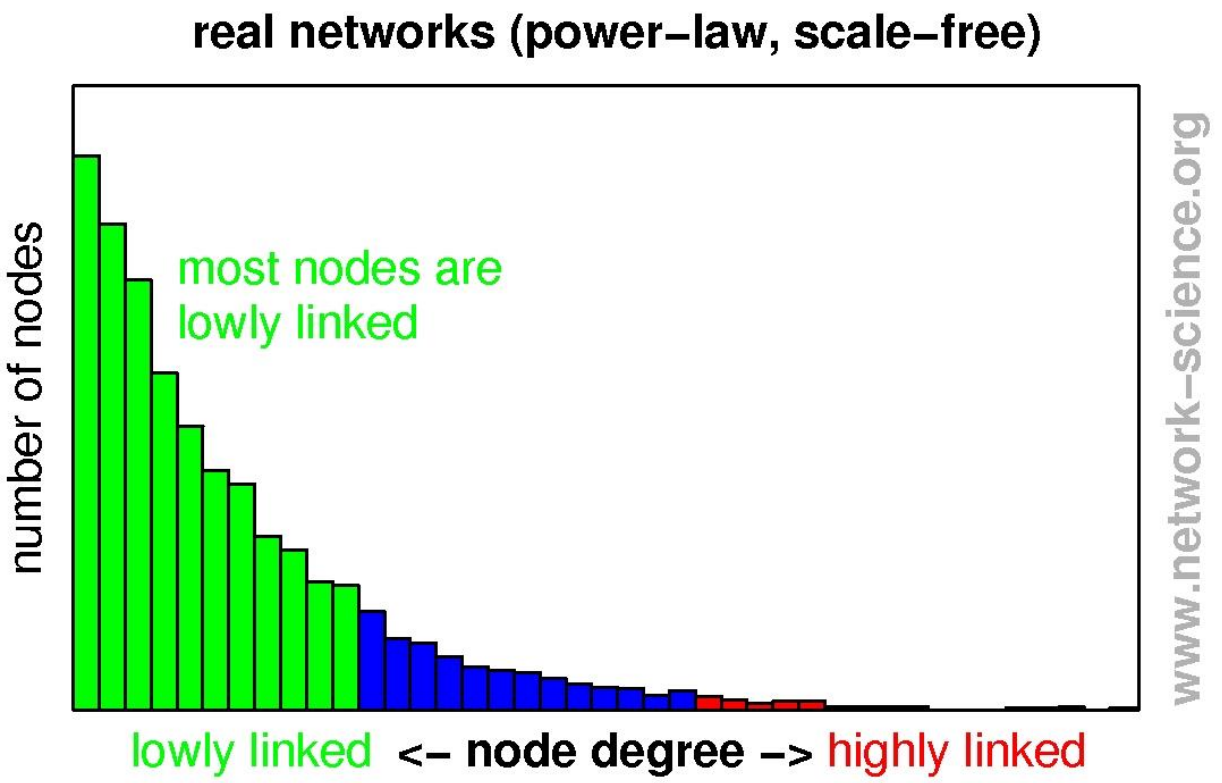
Find people who had no interaction with "Cecil" through any comments, neither replying nor receiving a reply.

The result is „Alice".

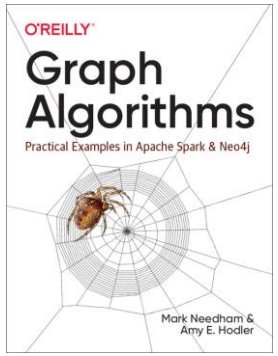
Typical execution times are high.

GRAPH ANALTYICS: NETWORK SCIENCE

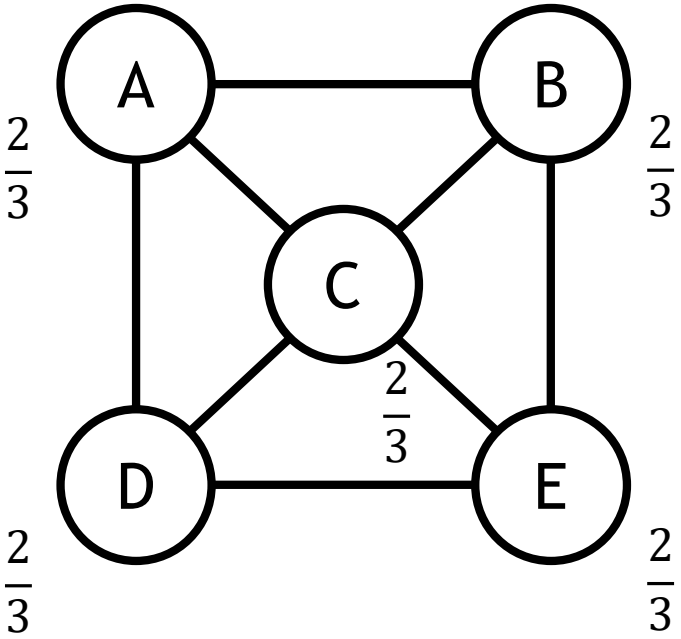




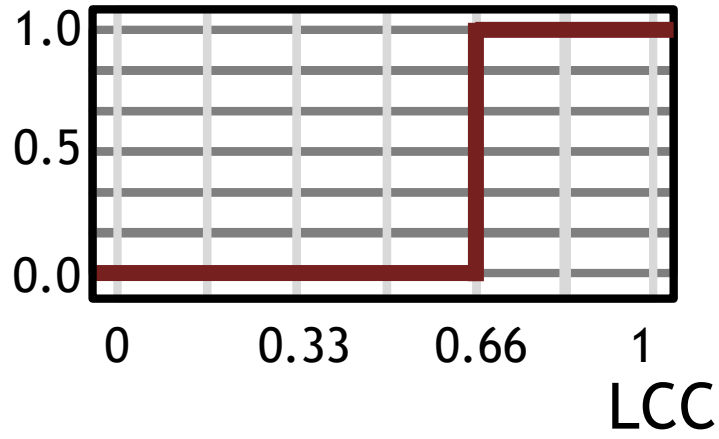
- Studies the *structure of graphs*
- Pioneered by László Barabási-Albert et al.
- Degree distributions, clustering coefficient, etc.



LOCAL CLUSTERING COEFFICIENT

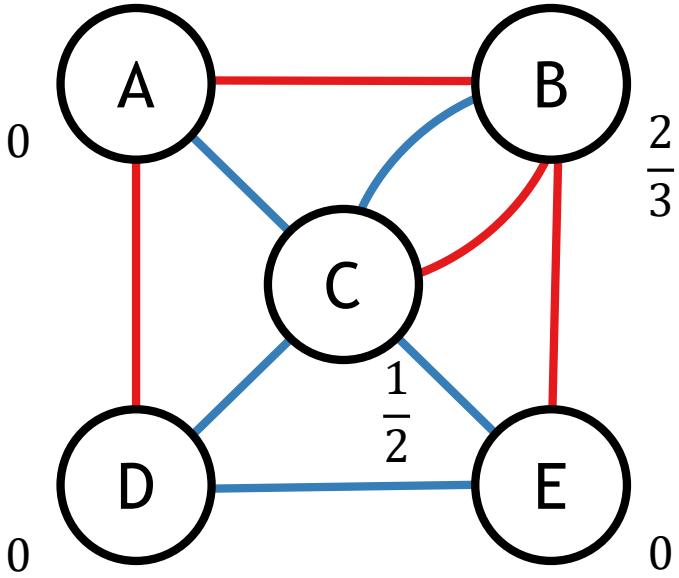


$$\mathrm{LCC}(v) = \frac{\text{triangles through } v}{\text{triples centered at } v}$$

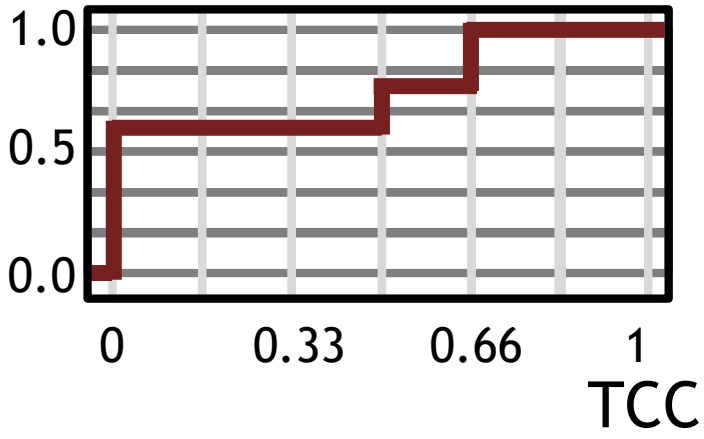


The empirical cumulative distribution function does not present much useful information in this case.

TYPED CLUSTERING COEFFICIENT



$$\mathrm{TCC}(v) = \frac{\text{v} + \text{v}}{\text{v} + \text{v}}$$

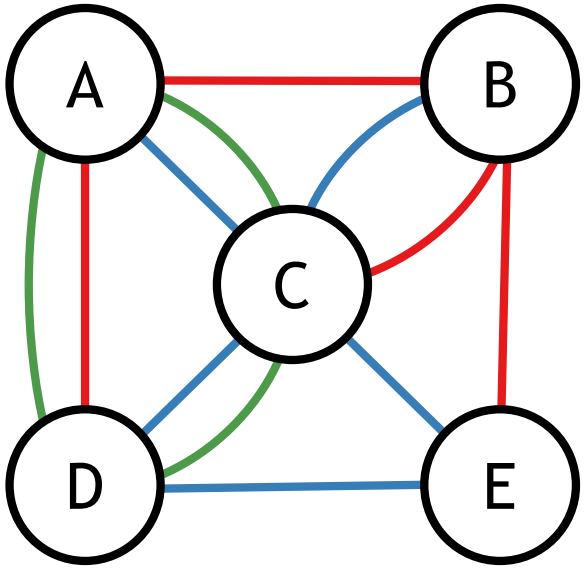


More information

High combinatorial complexity:

- t types $\rightarrow t \times (t-1)$ triangles
- $\mathcal{O}(t^2)$ steps

TYPED CLUSTERING COEFFICIENT



- Business partners
- Friends
- Family member

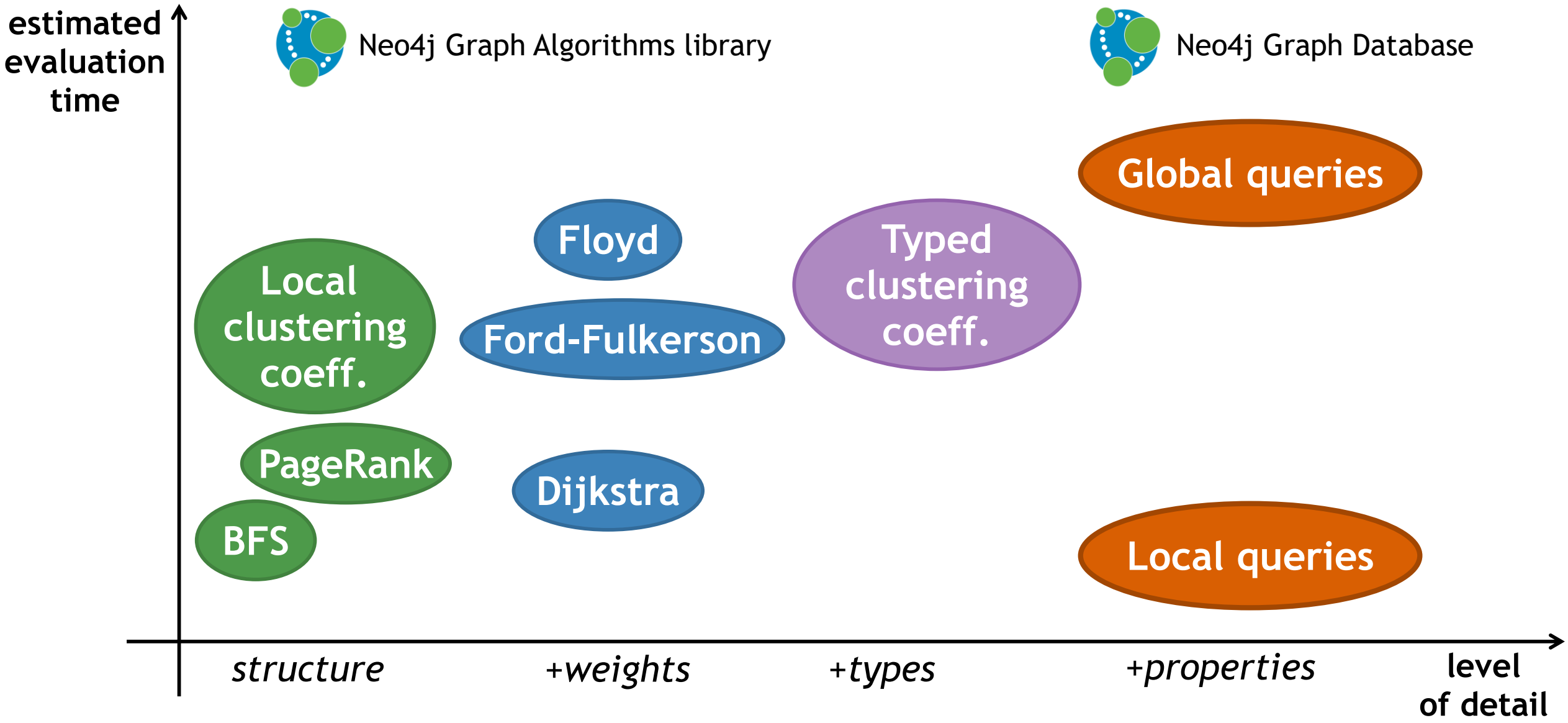
3 types → 6 triangles

$$\mathrm{TCC}(v) = \frac{\triangle_v + \triangle_v + \triangle_v + \triangle_v + \triangle_v + \triangle_v}{\wedge_v + \wedge_v + \wedge_v}$$

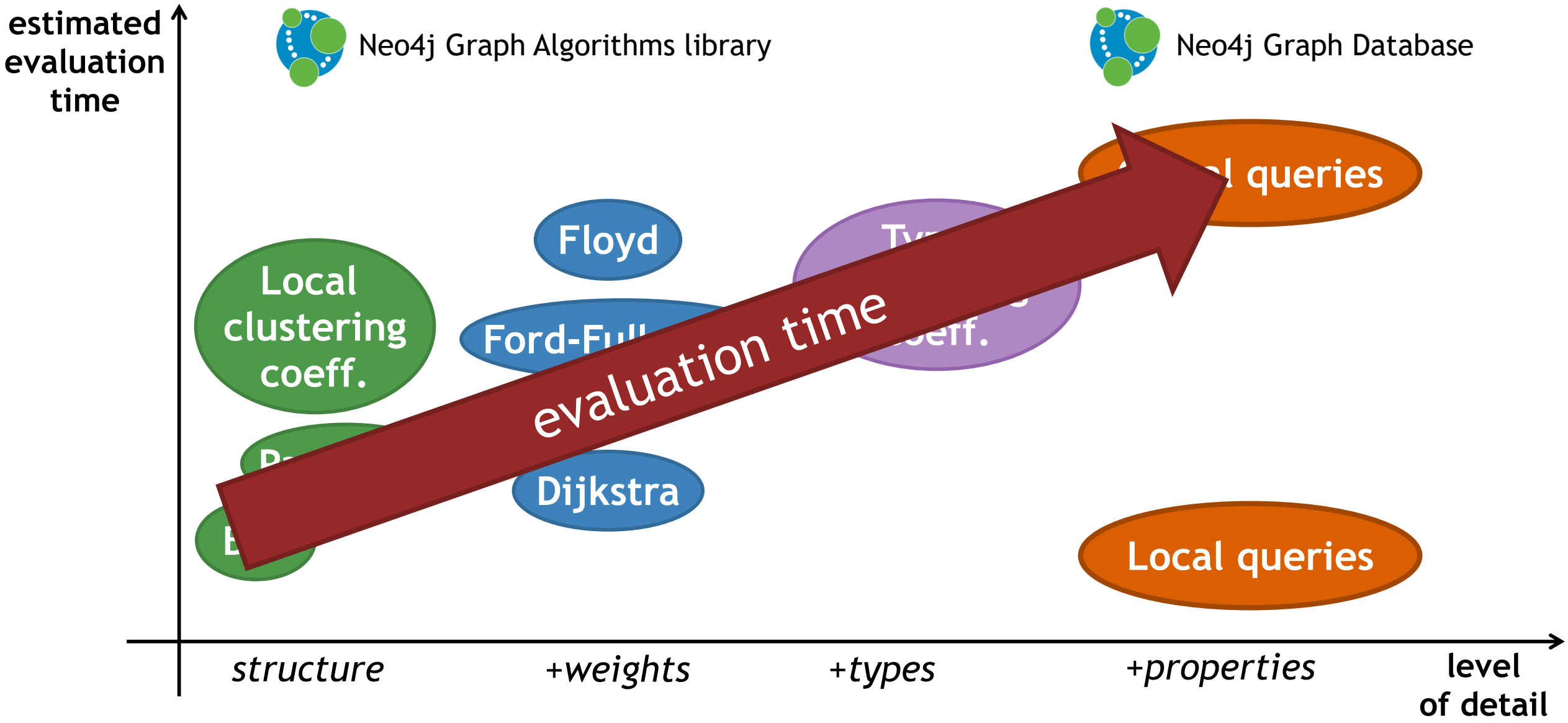
F. Battiston et al.:
Structural measures for multiplex networks,
Physical Review E, 2014

Petra Várhegyi:
Multidimensional Graph Analytics,
Master's thesis, 2018

GRAPH PROCESSING TECHNIQUES AND LANGUAGES



GRAPH PROCESSING TECHNIQUES AND LANGUAGES



Graph processing tools and challenges

GRAPH PROCESSING CHALLENGES / STRUCTURE

connectedness — the “curse of connectedness”

computer architectures — data structures contemporary computer architectures are good at processing are linear and simple hierarchical structures, such as *Lists*, *Stacks*, or *Trees*

caching and parallelization — a massive amount of random data access is required […] poor performance since the CPU cache is not in effect for most of the time. […] parallelism is difficult to extract because of the unstructured nature of graphs.

B. Shao, Y. Li, H. Wang, H. Xia (Microsoft Research):
Trinity Graph Engine and its Applications,
IEEE Data Engineering Bulleting 2017

GRAPH PROCESSING CHALLENGES / PROPERTIES

topology

existing graph query methods […] focus on the topological structure of graphs and few have considered attributed graphs.

properties

applications of large graph databases would involve querying the graph data (attributes) in addition to the graph topology.

complex optimization

answering queries that involve predicates on the attributes of the graphs in addition to the topological structure […] makes evaluation and optimization more complex.


G-SPARQL: A Hybrid Engine for Querying Large Attributed Graphs

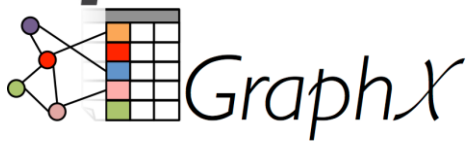
S. Sakr, S. Elnikety, Y. He (Microsoft Research):
G-SPARQL: A Hybrid Engine for Querying Large Attributed Graphs,
CIKM 2012

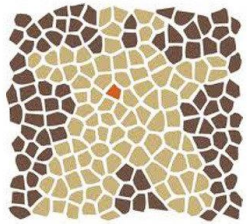
GRAPH PROCESSING TOOLS

Currently, there is a strong distinction between graph query and analytical tools – this might change in the future.

graph queries

graph analytics





Gelly

LynxKite

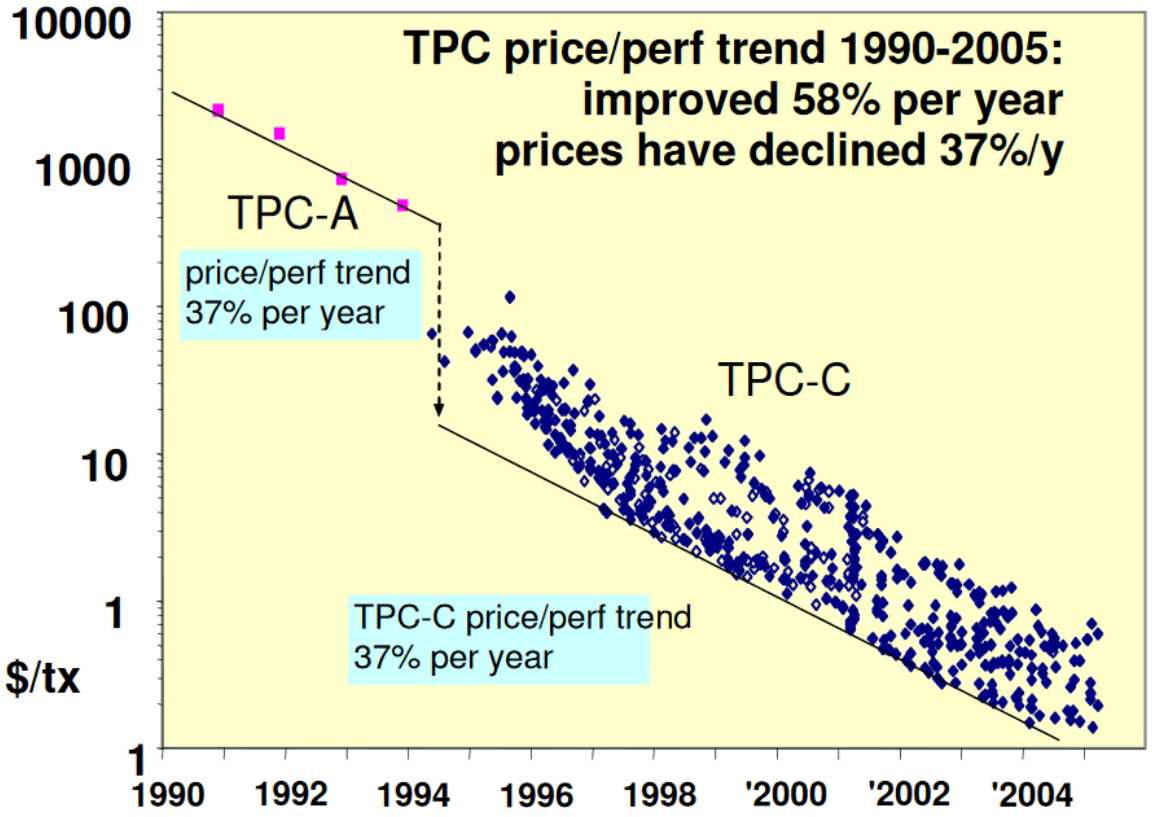
János Szendi-Varga (GraphAware):
Graph Technology Landscape 2019

Neo4j Graph Algorithms library

Benchmarks:
Defining a common understanding

TRANSACTION PROCESSING PERFORMANCE COUNCIL (1988-)

Many standard specifications for benchmarking certain aspects of relational DBs



- Transaction Processing - OLTP
  - TPC-C
  - TPC-E
- Decision Support
  - TPC-H
  - TPC-DS
  - TPC-DI
- Virtualization
  - TPC-VMS
  - TPCx-V
  - TPCx-HCI
- Big Data
  - TPCx-HS V1
  - TPCx-HS V2
  - TPCx-BB
- IoT
  - TPCx-IoT
- Common Specifications
  - TPC-Energy
  - TPC-Pricing

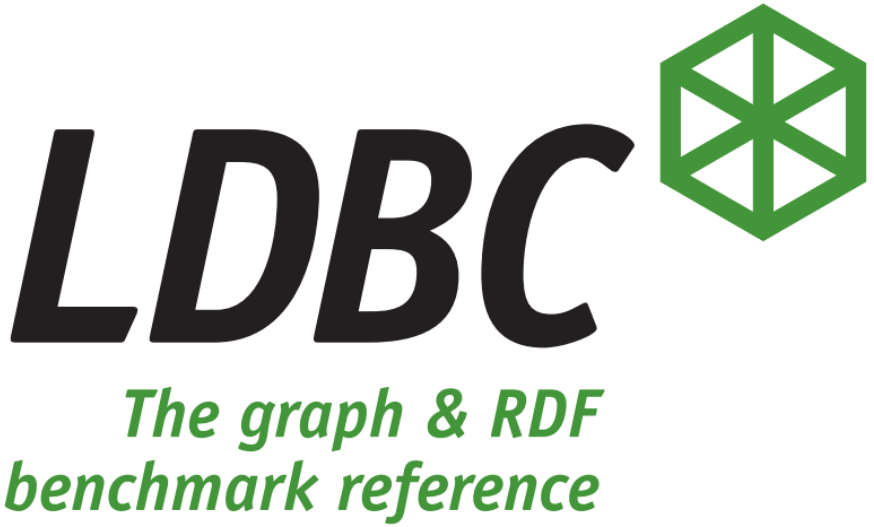
- Top Ten
  - Most Recent Ten Published
  - All Results
  - Advanced Sort
- By Perf.
- By Price/Perf.

TPC™

LINKED DATA BENCHMARK COUNCIL (2012-)

LDBC is a non-profit organization dedicated to establishing benchmarks, benchmark practices and benchmark results for graph data management software.

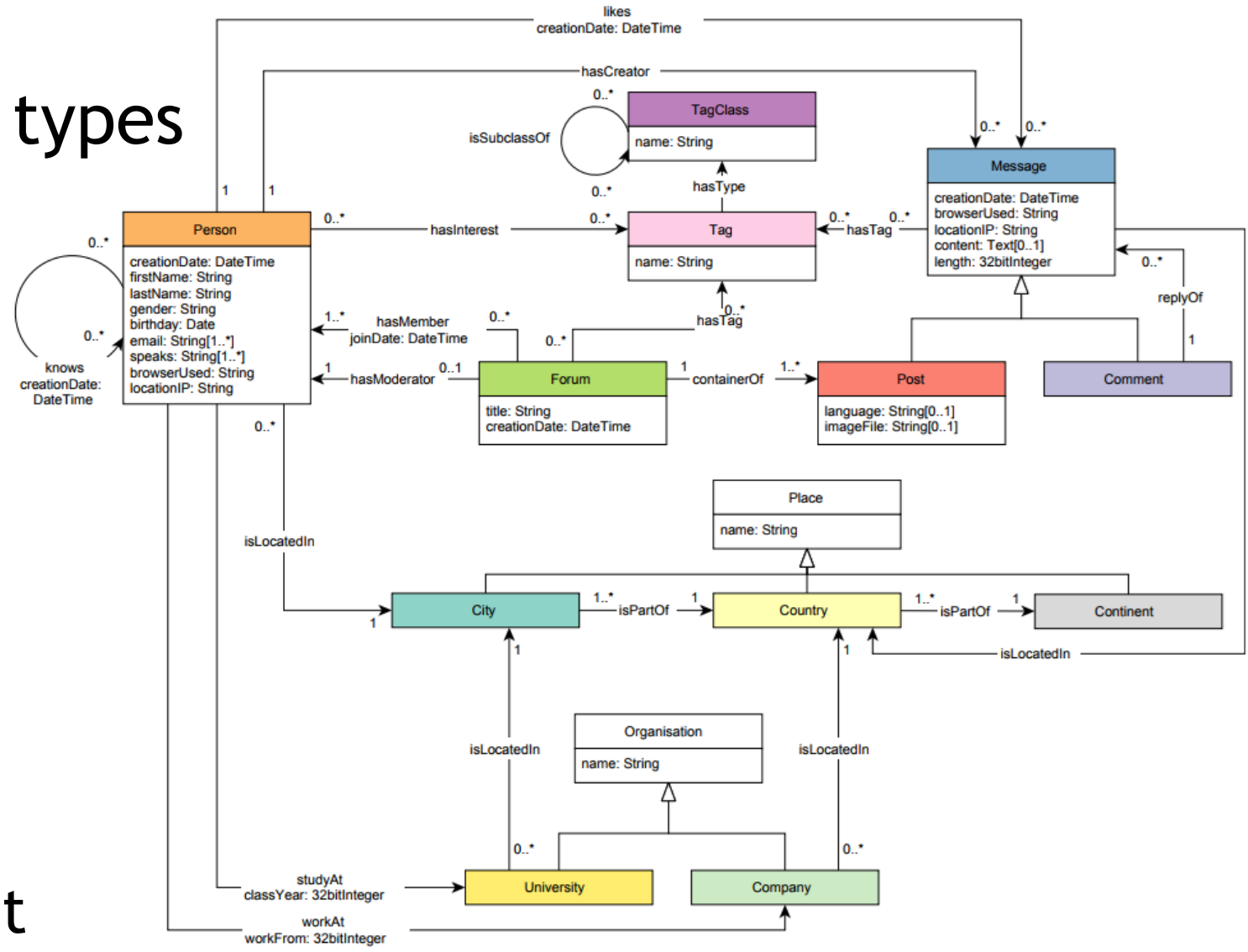
LDBC's Social Network Benchmark is an industrial and academic initiative, formed by principal actors in the field of graph-like data management.



LDBC SOCIAL NETWORK BENCHMARK

Complex graph schema

14 node types, many edge types

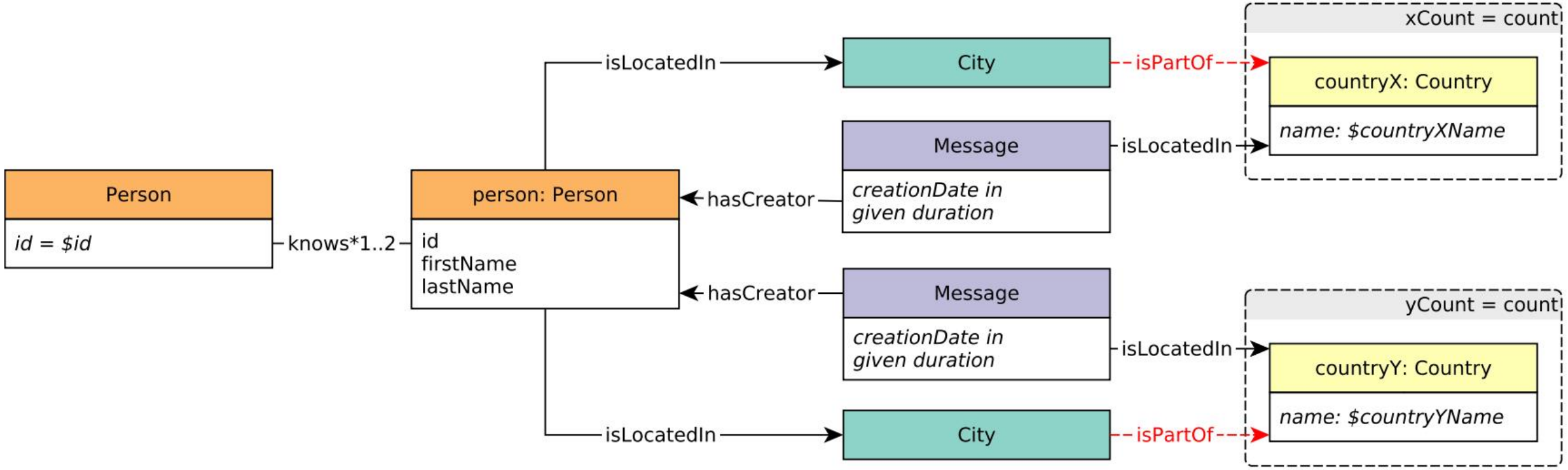


Subgraphs

- Network of persons
- Arbitrary depth trees
  - Comments
  - TagClasses
- Fixed depth trees
  - City < Country < Continent

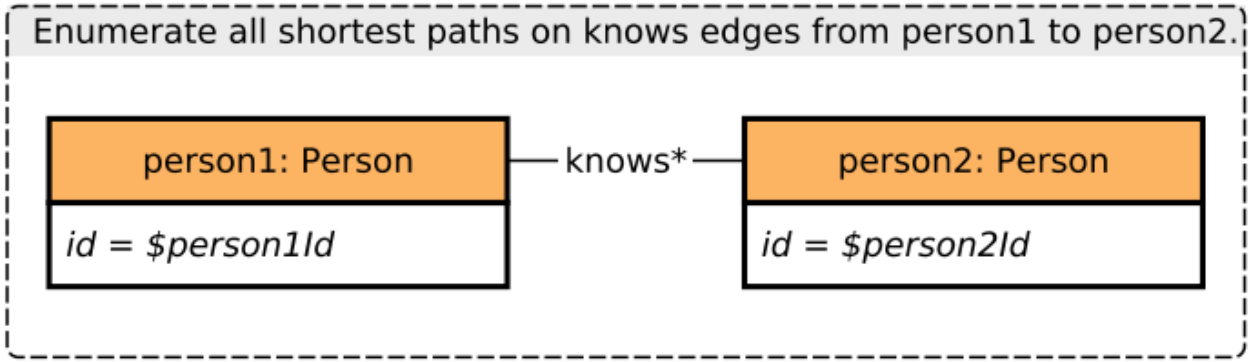
LDBC INTERACTIVE Q3

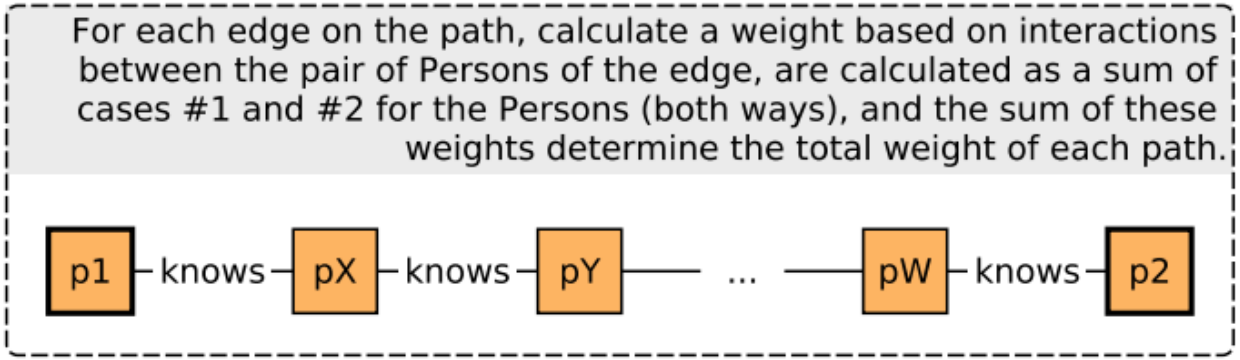
Friends and **friends of friends** that have been to countries X and Y

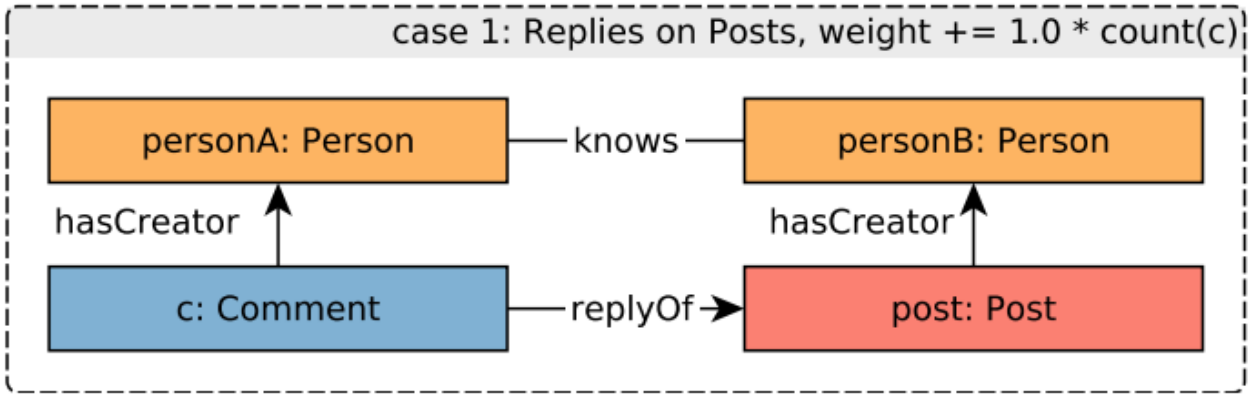


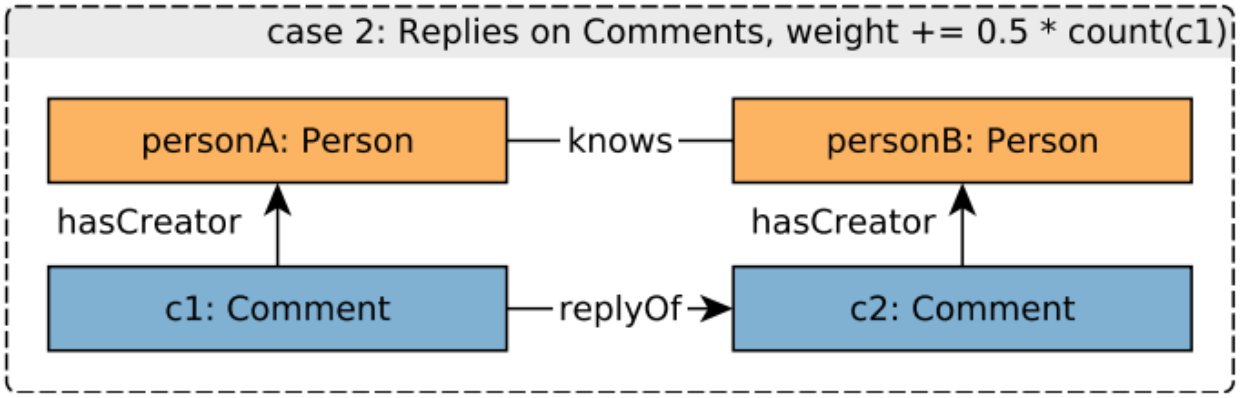
LDBC INTERACTIVE Q14

Trusted connection paths

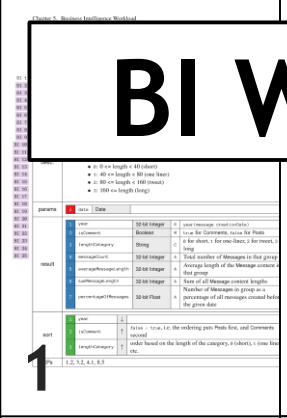


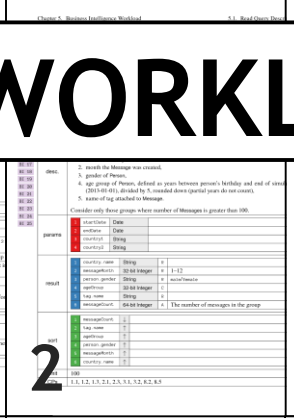


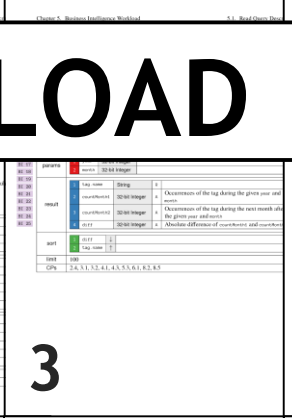


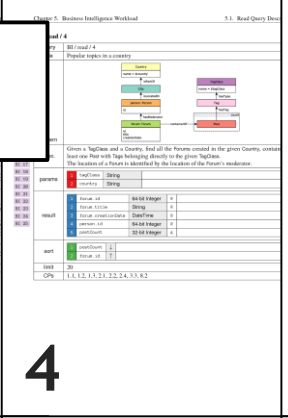


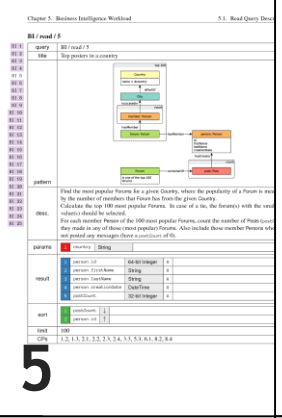
BI WORKLOAD

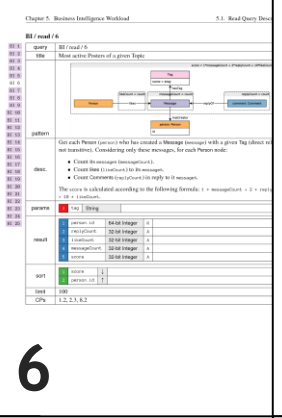

1

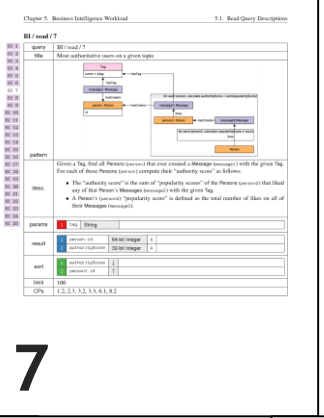

2


3


4

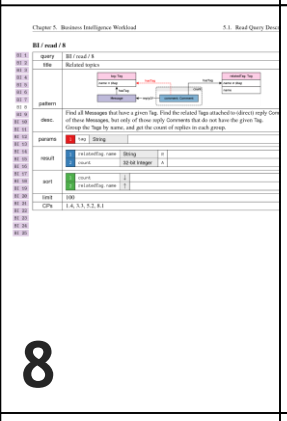

5

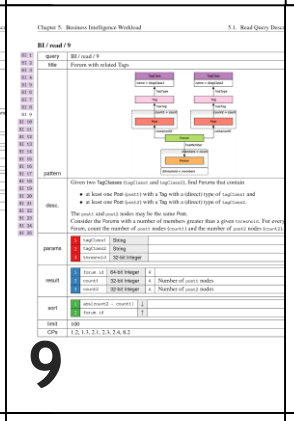

6

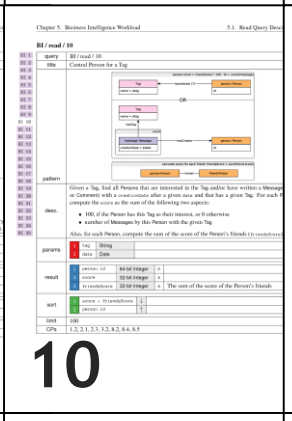

7

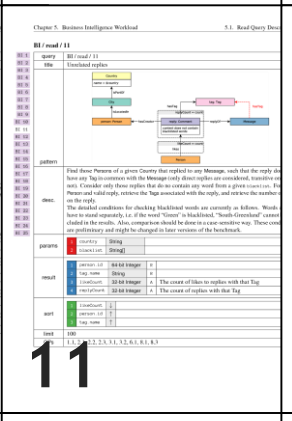
LDBC

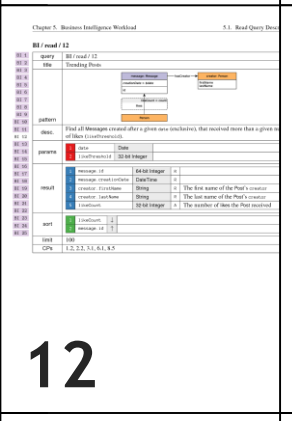
The graph & RDF benchmark reference

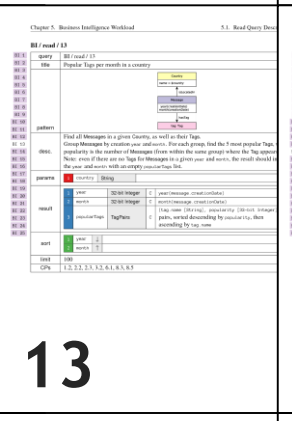

8

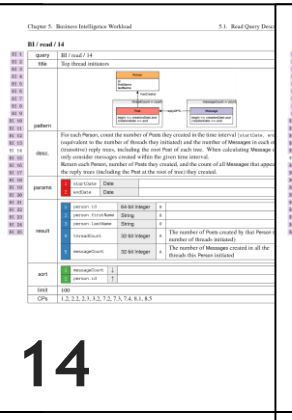

9

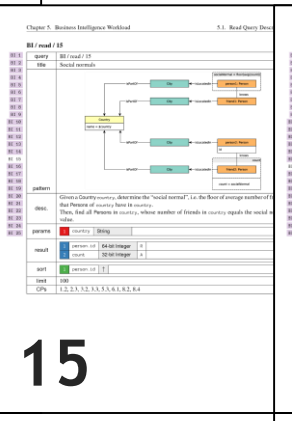

10

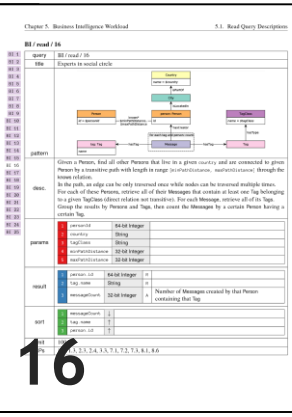

11

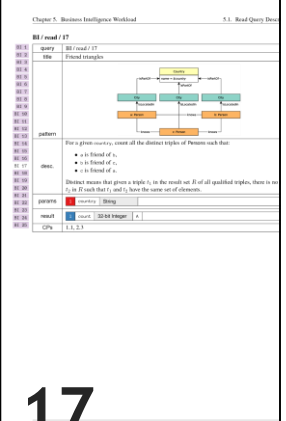

12

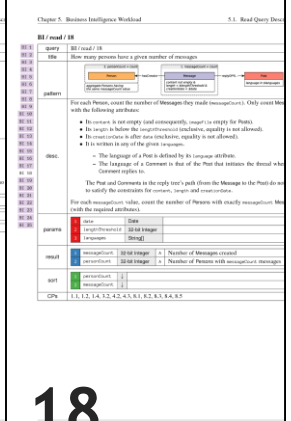

13

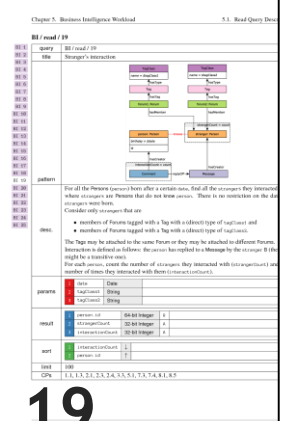

14

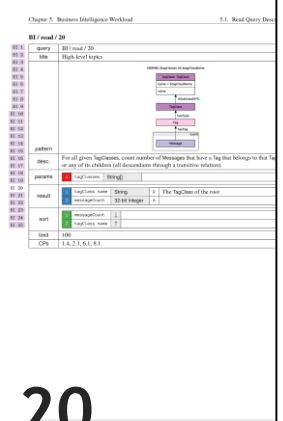

15

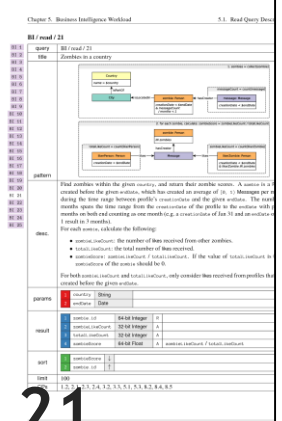

16

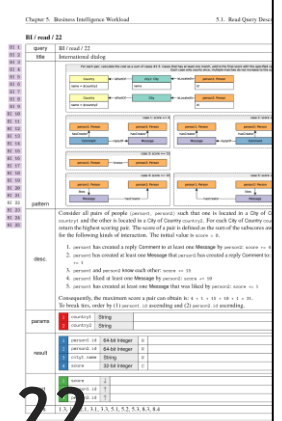

17

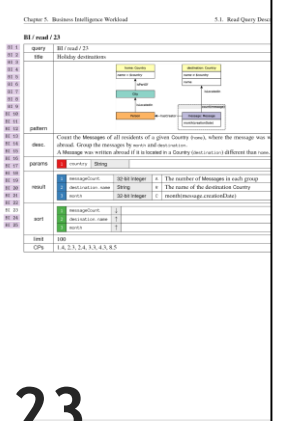

18

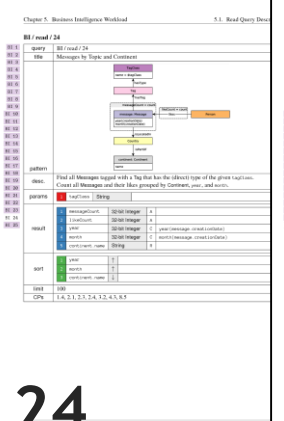

19

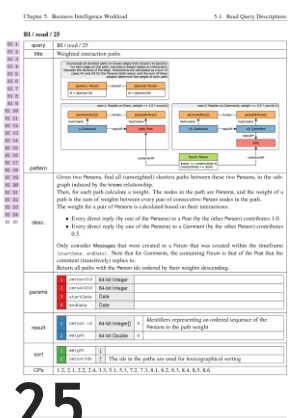

20


21


22


23

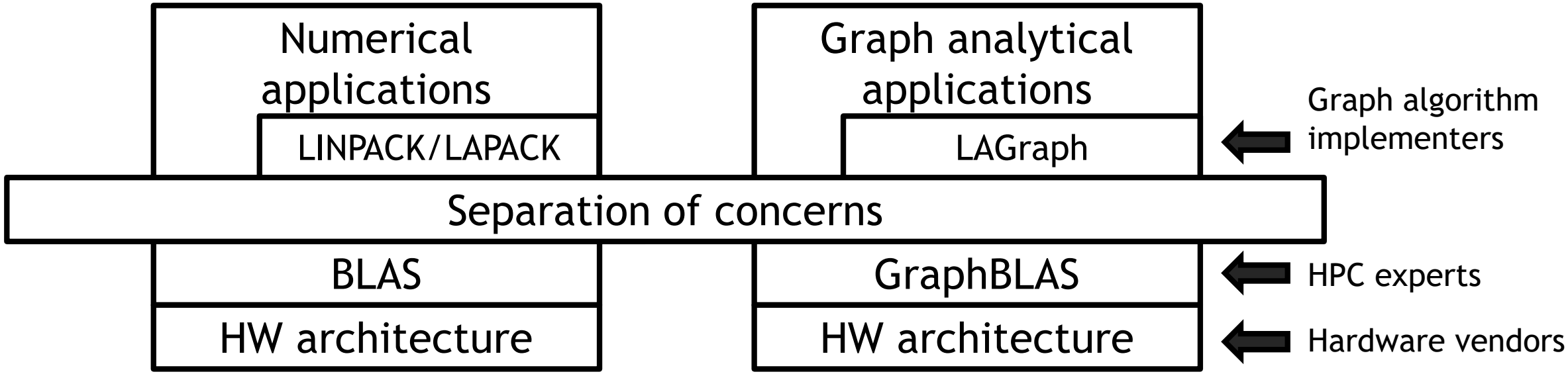

24

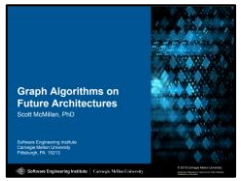

25

GraphBLAS:
A unified theory built on linear algebra

THE GRAPHBLAS APPROACH

- GraphBLAS is an effort to define standard building blocks for graph algorithms in the language of linear algebra
- 1979: BLAS (Basic Linear Algebra Subprograms)
- 2013: GraphBLAS
- Key idea: *separation of concerns*





S. McMillan: Research review @ CMU, 2015
Graph algorithms on future architectures

Tim Mattson et al.: LAGraph, GrAPL @ IPDPS 2019

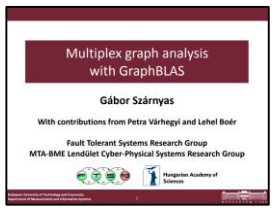
PARALLELIZATION ON SKEWED DISTRIBUTIONS

Using multiple processing units require **load balancing.**

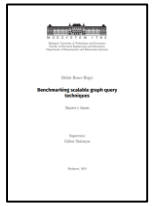
Very difficult to implement for real graphs.

This work is in progress and improvements are expected.

Dataset	Algorithm	Single thread	Multi thread	Ratio
graph500-22	LCC	3083.494	261.193	11.80
datagen-7_9-fb	LCC	938.139	213.783	4.38
datagen-7_6-fb	LCC	431.211	101.173	4.26
datagen-7_5-fb	LCC	336.583	80.637	4.17
datagen-7_7-zf	LCC	183.948	138.022	1.33
datagen-7_8-zf	LCC	234.518	176.558	1.32



Gábor Szárnyas: *Multiplex graph analytics with GraphBLAS*, FOSDEM 2019



Bálint Hegyi: *Benchmarking scalable graph query techniques*, Master's thesis, 2019

Summary

SUMMARY: CHALLENGES IN GRAPH PROCESSING

No consensus on a unifying theory:

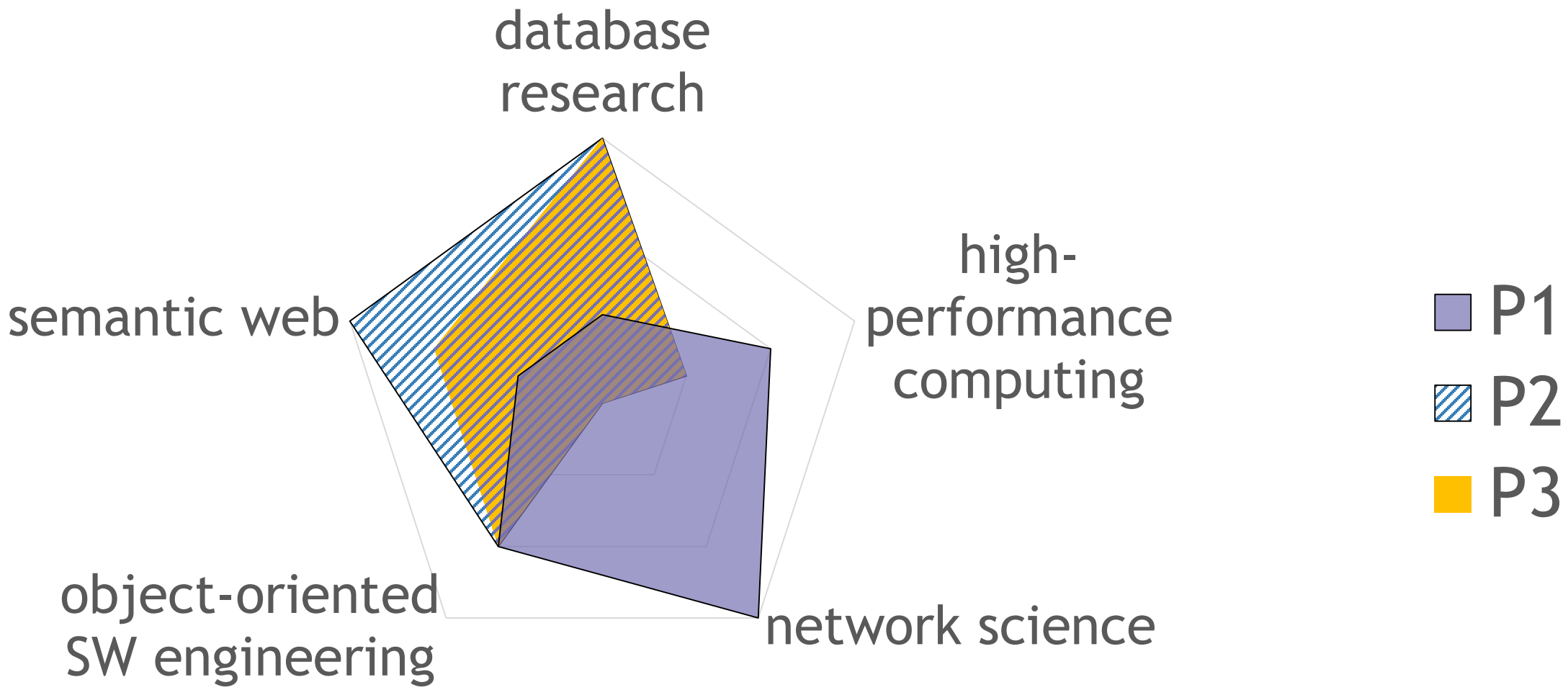
- Relational algebra?
- Linear algebra?

Performance:

- Many random access operations
- Difficult to cache
- Difficult to parallelize
- Handling properties introduces even more complexity

Many open research and implementation challenges.

CONTRIBUTIONS IN MY PHD DISSERTATION



Gábor Szárnyas:
Query, Analysis, and Benchmarking Techniques for Evolving Property Graphs of Software Systems,
PhD dissertation, 2019