



Graph Databases after 15 Years – Where Are They Headed?

Gábor Szárnyas

Data Analytics devroom | FOSDEM | 2025-02-01

About me

2012–2019
MSc & PhD



graphs

2020–2023
postdoc



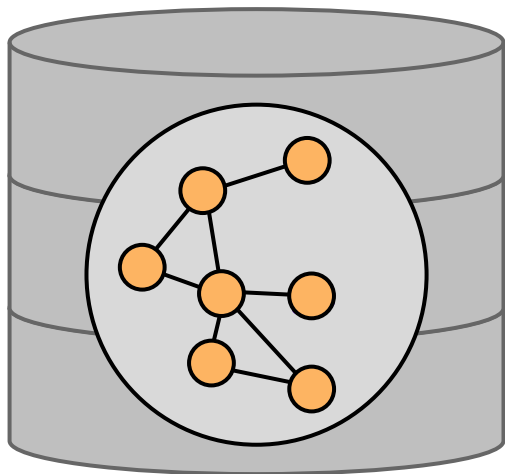
graphs

2023–
devrel



DuckDB Labs

not graphs



1

graph databases
in 15 minutes

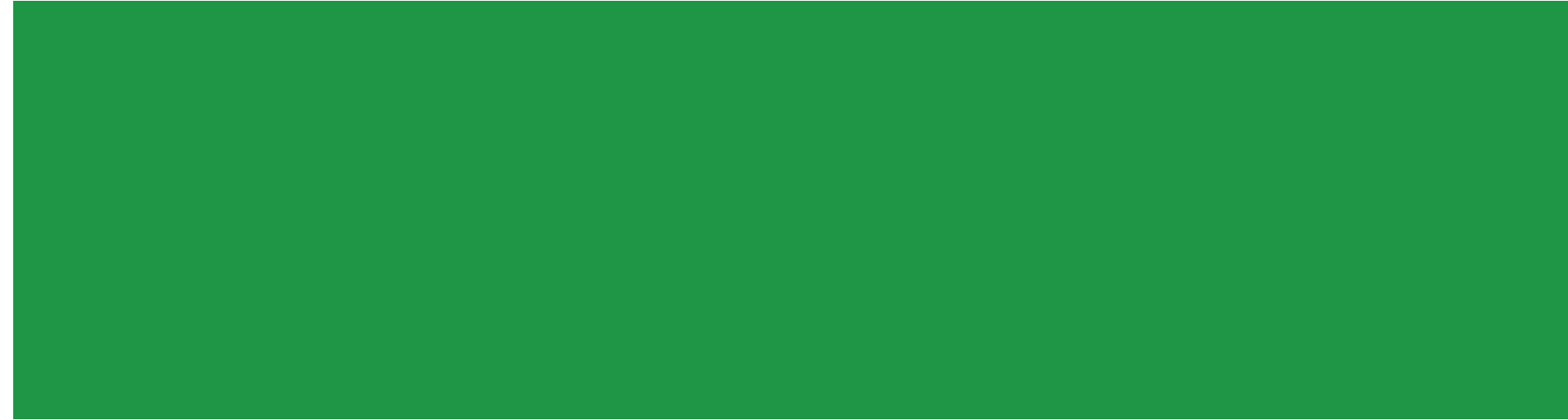
2

new research
– not just for graphs!

3

pointers
for benchmarks

Running example

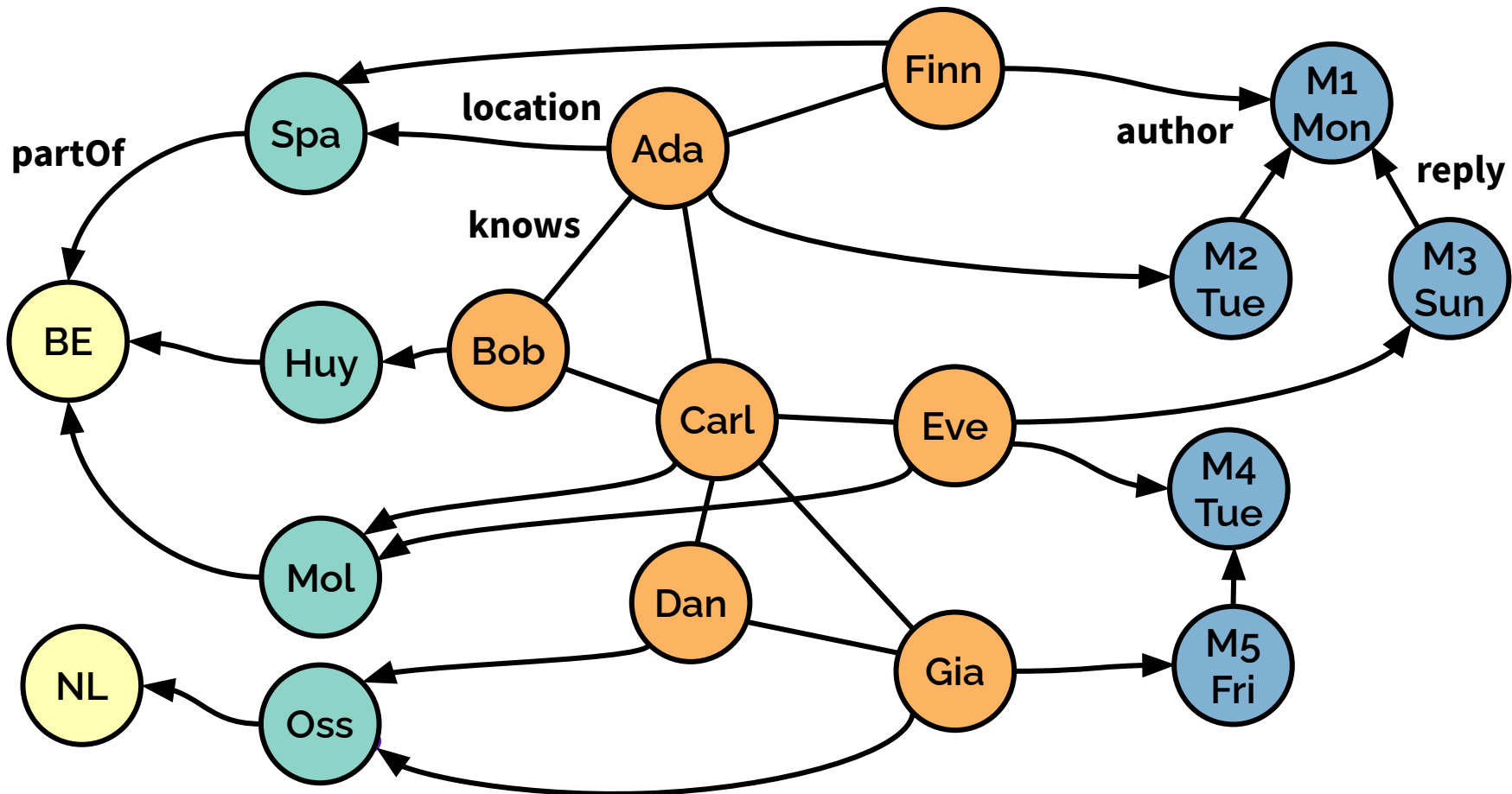


Country

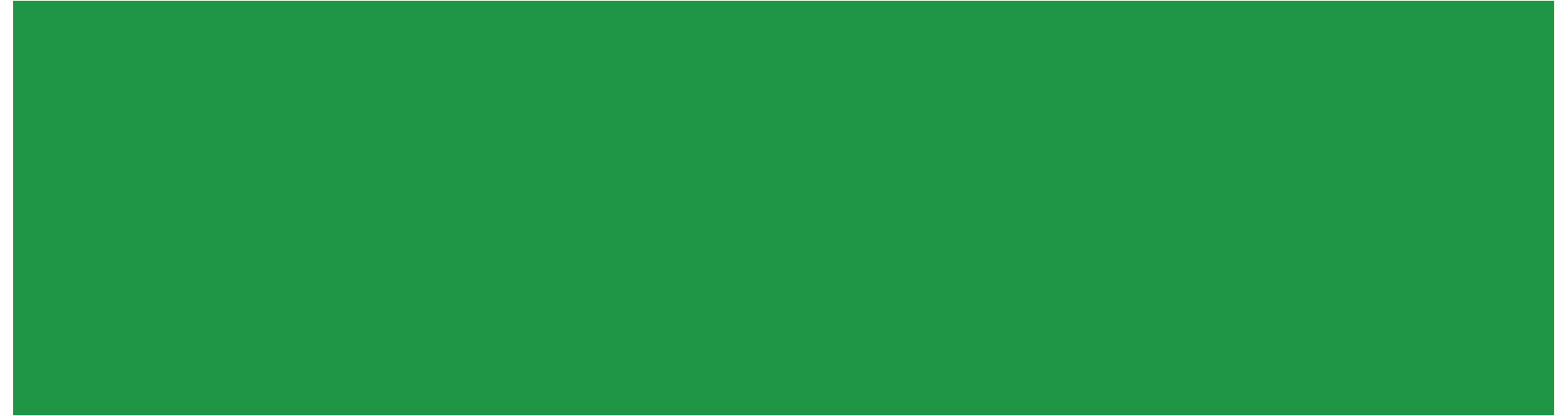
City

Person

Message



Graph databases



**DO NOT
DISTURB**



Graph
Construction Zone
No Joins
beyond this point

Fast data processing

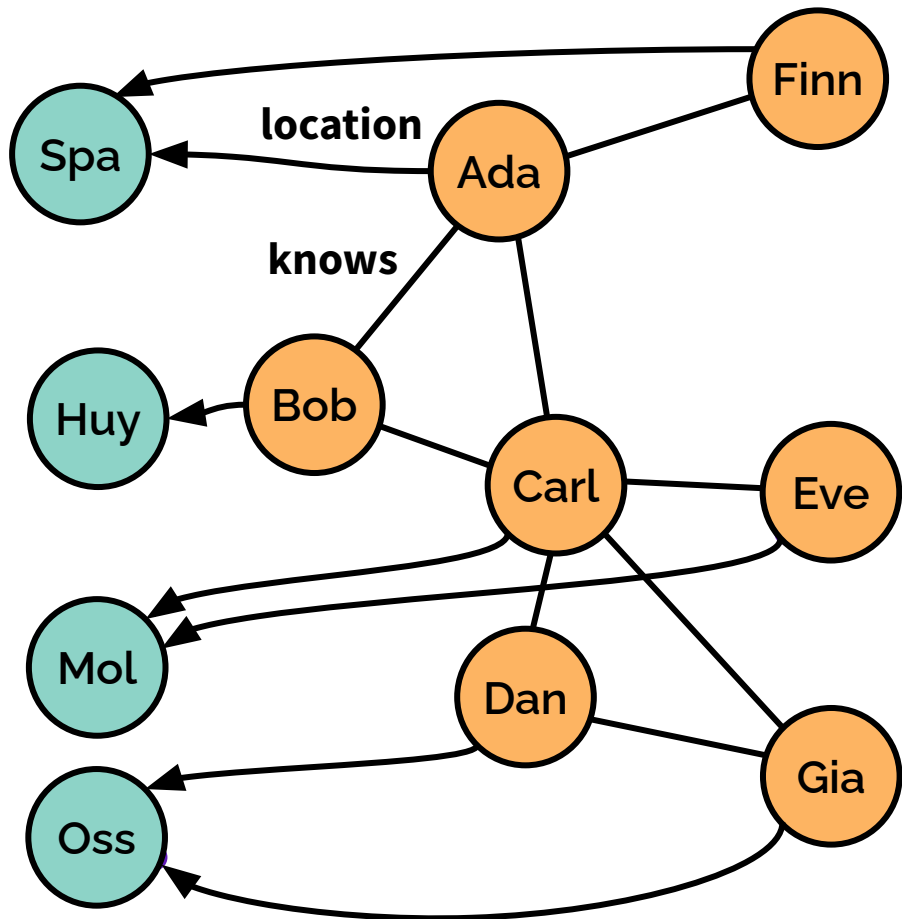
No expensive run-time JOINS.

Join those who
left **JOINS**
behind

Relational databases can't join?



Where do Dan's friends live?



knows

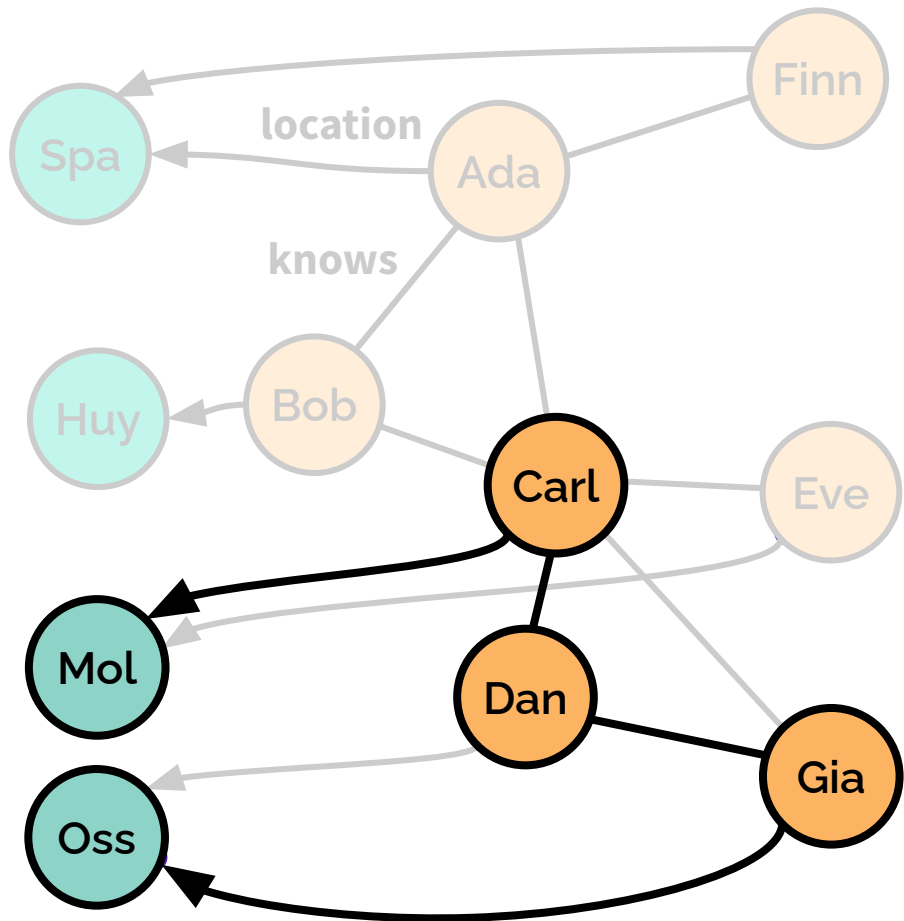
p1	p2
Ada	Bob
Ada	Carl
Ada	Finn
Bob	Huy
Carl	Dan
Carl	Eve
Carl	Gia
Dan	Gia

location

person	city
Ada	Spa
Bob	Huy
Carl	Mol
Dan	Oss
Eve	Mol
Finn	Spa
Gia	Oss

⋈

Where do Dan's friends live?



knows

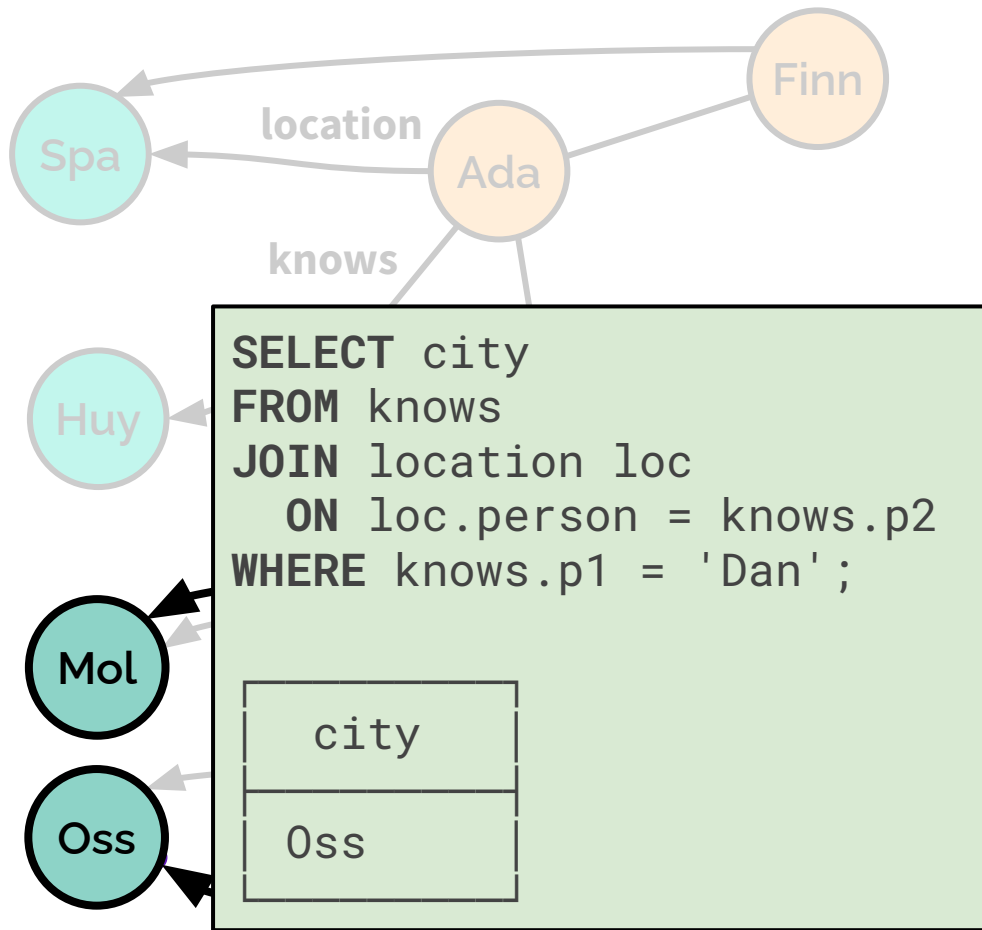
p1	p2
Ada	Bob
Ada	Carl
Ada	Finn
Bob	Carl
Carl	Dan
Carl	Eve
Carl	Gia
Dan	Gia

location

person	city
Ada	Spa
Bob	Huy
Carl	Mol
Dan	Oss
Eve	Mol
Finn	Spa
Gia	Oss

⋈

Where do Dan's friends live?



knows

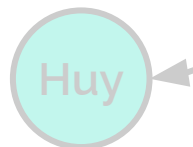
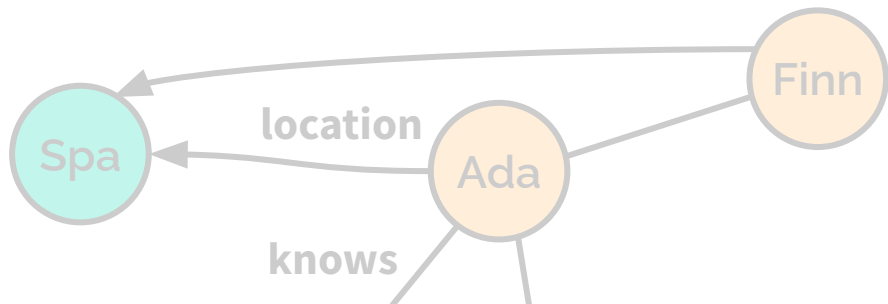
p1	p2
Ada	Bob
Ada	Carl
Ada	Finn
Bob	Carl
Carl	Dan
Carl	Eve
Carl	Gia
Dan	Gia

location

person	city
Ada	Spa
Bob	Huy
Carl	Mol
Dan	Oss
Eve	Mol
Finn	Spa
Gia	Oss

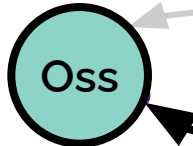
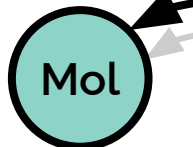
⋈

Where do Dan's friends live



```
SELECT city
FROM knows
JOIN location loc
  ON loc.person = knows.p2
WHERE knows.p1 = 'Dan';
```

city
Oss

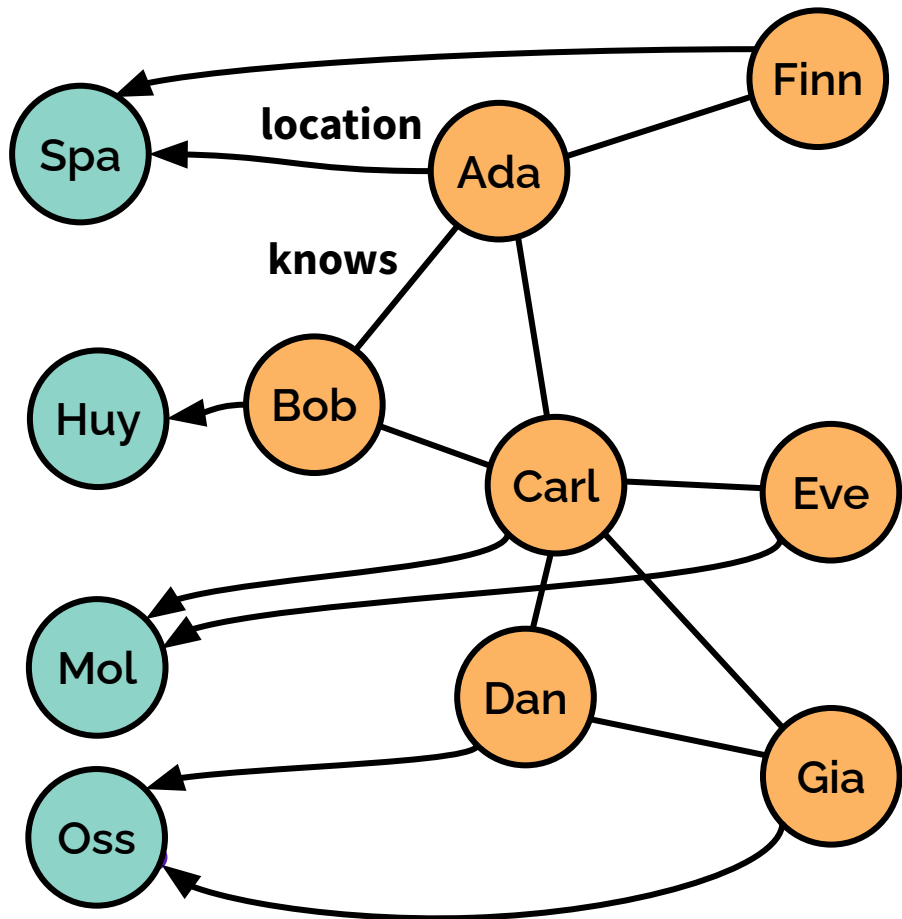


```
CREATE VIEW knows_undir AS
SELECT p1, p2 FROM knows
UNION ALL
SELECT p2, p1 FROM knows;
```

```
SELECT city
FROM knows_undir ku
JOIN location loc
  ON loc.person = ku.p2
WHERE ku.p1 = 'Dan';
```

city varchar
Mol Oss

Where do Dan's friends live?



Cypher query language

```
MATCH (p1 {name: 'Dan'})  
      -[:knows]-(p2)  
      -[:location]->(c)  
RETURN c.name;
```

Two advantages:

- concise and readable joins
- elegant handling of bidirectional edges

Most graph databases have some special syntax sugar for joins.

transactional

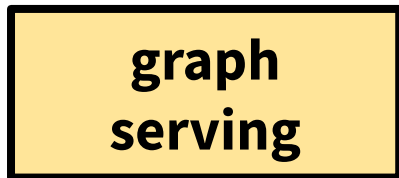
analytical



all three categories address join problems

transactional

analytical



repeated joins =
n + 1 query problem

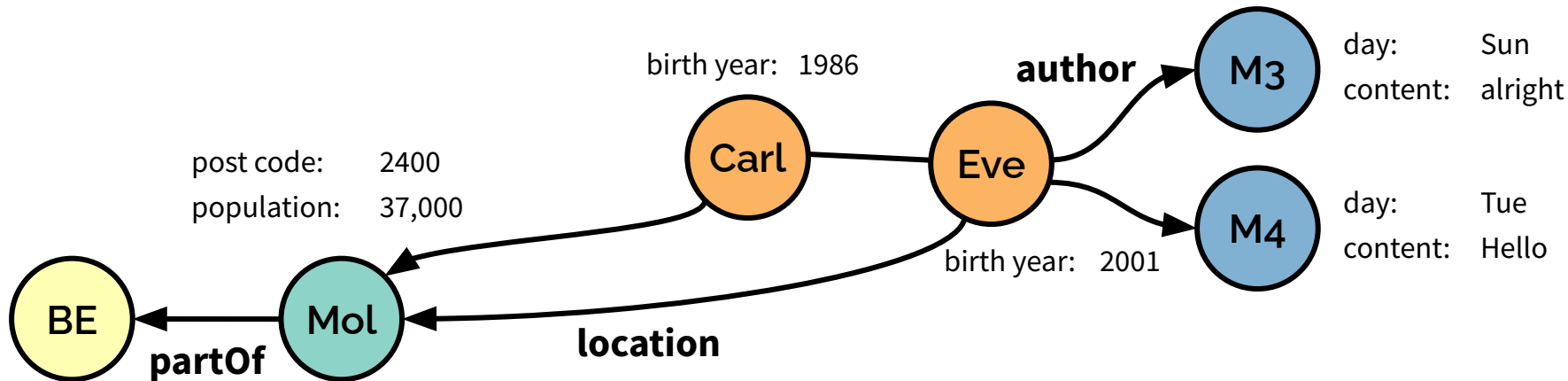


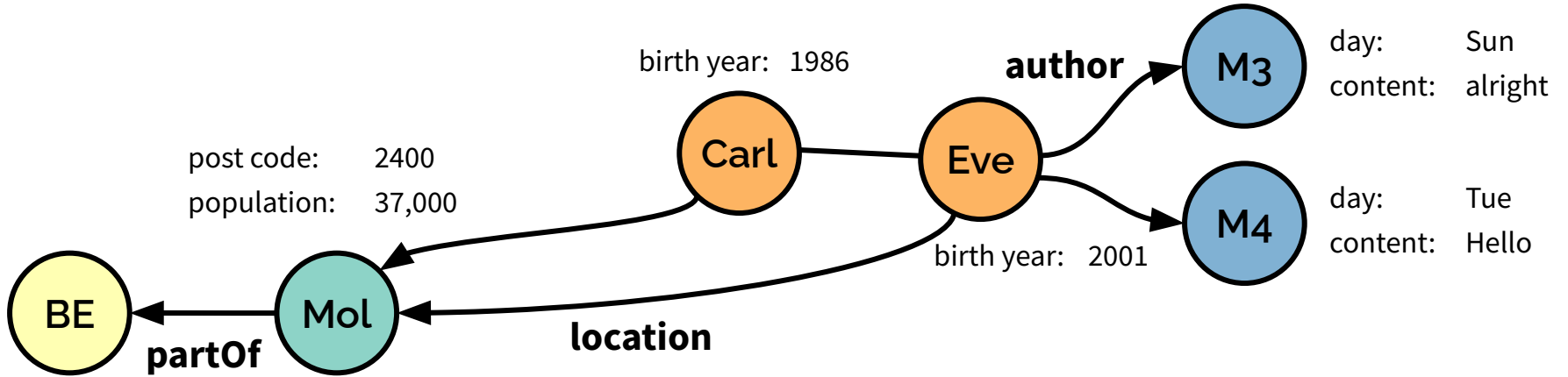
Country

City

Person

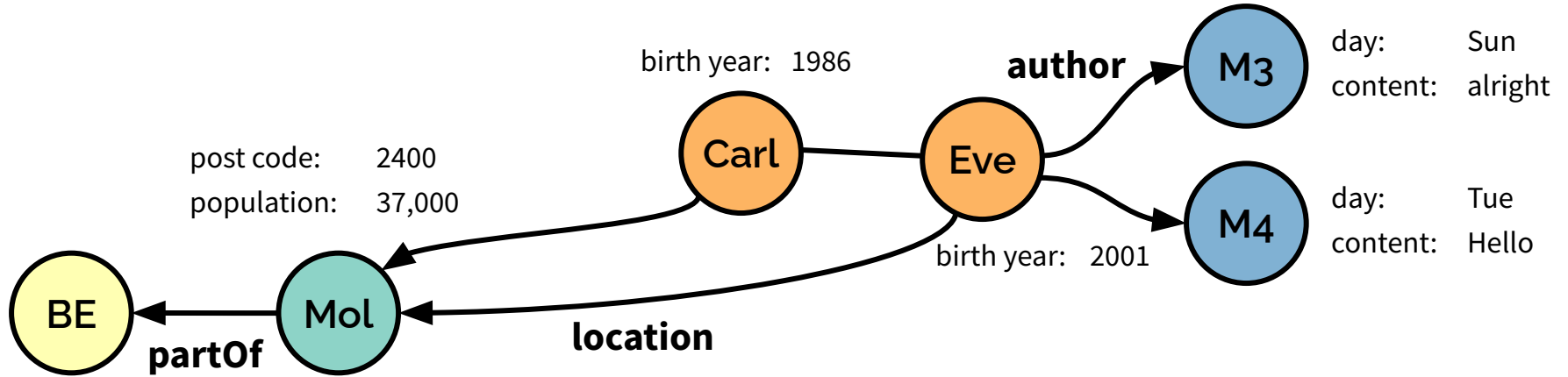
Message





partOf ⋈ **City** ⋈ **location** ⋈ **Person** ⋈ **author** ⋈ **Message** (left joins + filtering)

city name	post code	population	person name	birth year	message id	day	comment
Mol	2400	37,000	Carl	1986	NULL	NULL	NULL
Mol	2400	37,000	Eve	2001	M3	Sun	alright
Mol	2400	37,000	Eve	2001	M4	Tue	Hello



partOf ⋈ **City** ⋈ **location** ⋈ **Person** ⋈ **author** ⋈ **Message** (left joins + filtering)

city name	post code	population	person name	birth year	message id	day	comment
Mol	2400	37,000	Carl	1986	NULL	NULL	NULL
Mol	2400	37,000	Eve	2001	M3	Sun	alright
Mol	2400	37,000	Eve	2001	M4	Tue	Hello

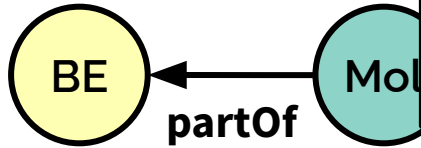
Country

City

Person

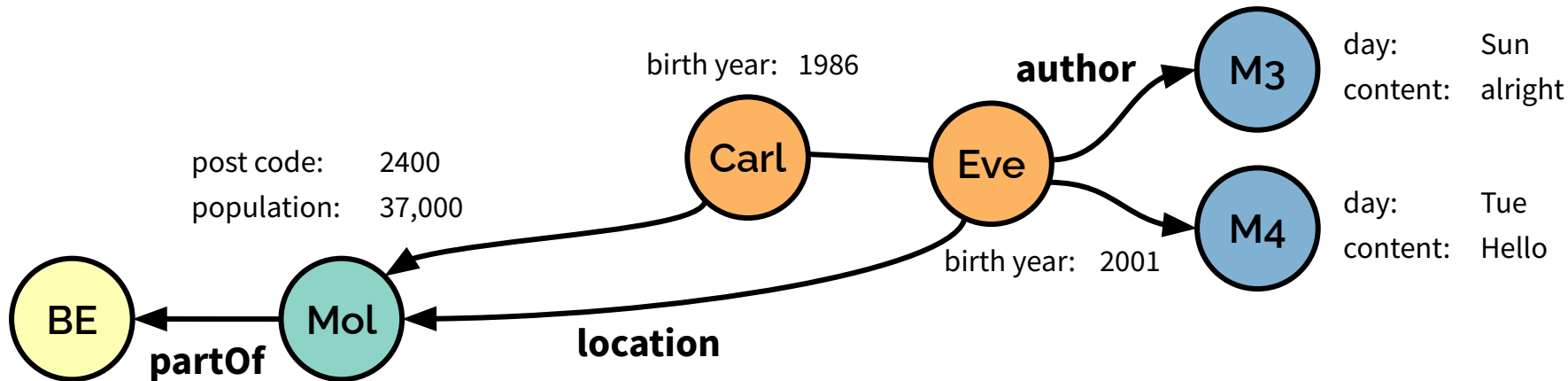
Message

**complex queries + overfetching +
client has to reconstruct the graph
from a table**



partOf ⋈ **City** ⋈ **location** ⋈ **Person** ⋈ **author** ⋈ **Message** (left joins + filtering)

city name	post code	population	person name	birth year	message id	day	comment
Mol	2400	37,000	Carl	1986	NULL	NULL	NULL
Mol	2400	37,000	Eve	2001	M3	Sun	alright
Mol	2400	37,000	Eve	2001	M4	Tue	Hello

Country**City****Person****Message**

city name	post code	population
Mol	2400	37,000

city name	person name	birth year
Mol	Carl	1986
Mol	Eve	2001

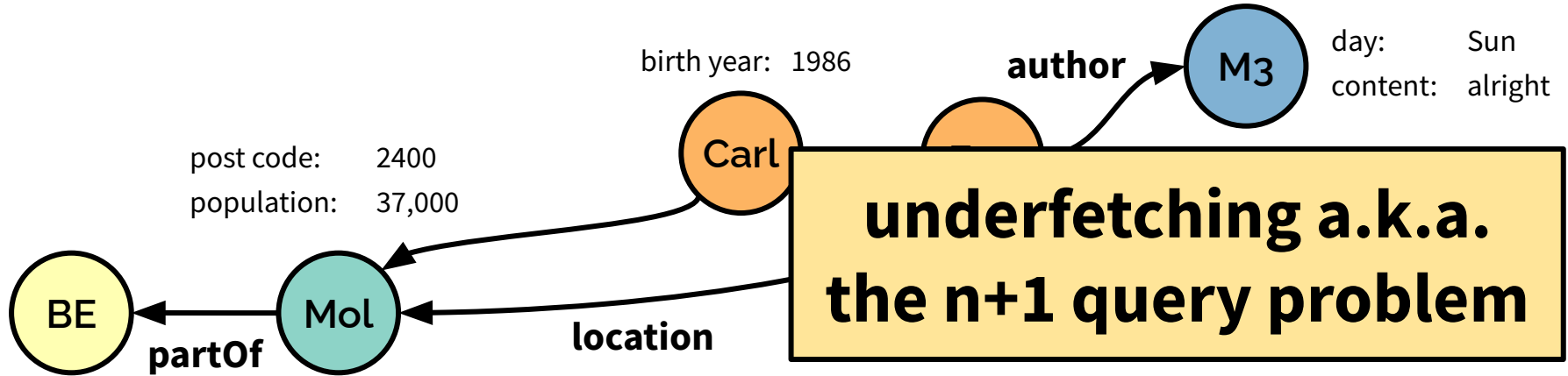
person name	message id	day	content
Eve	M3	Sun	alright
Eve	M4	Sun	Hello

Country

City

Person

Message



city name	post code	population
Mol	2400	37,000

city name	person name	birth year
Mol	Carl	1986
Mol	Eve	2001

person name	message id	day	content
Eve	M3	Sun	alright
Eve	M4	Sun	Hello

Why is querying a graph difficult?

Problems:

- *overfetching* is expensive, introduces redundancy
- *underfetching* is slow and leads to the n+1 query problem

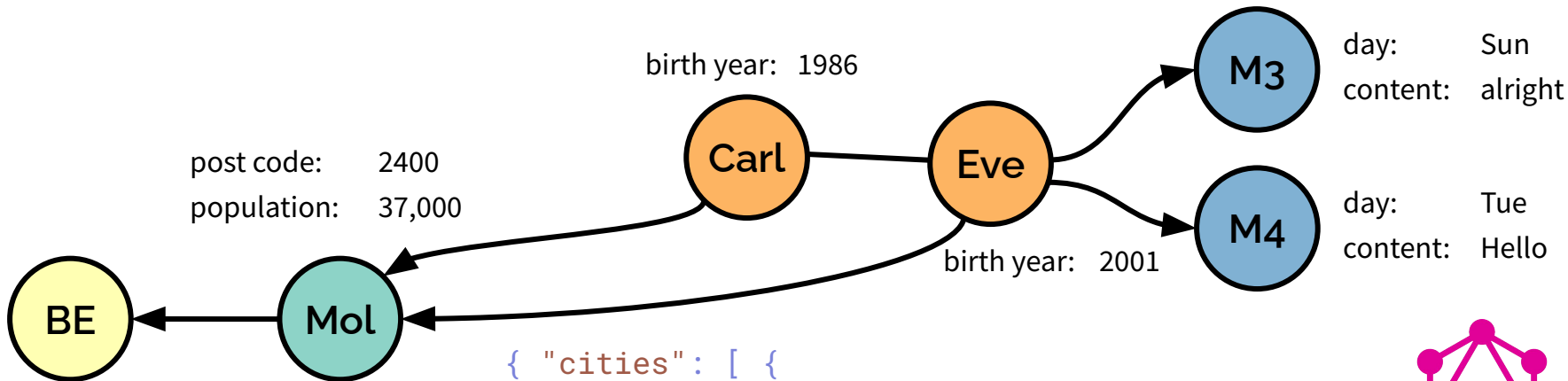
These are due to the **impedance mismatch from ORM** (object-relational mapping)

Country

City

Person

Message



```

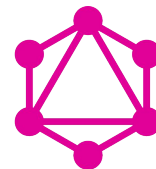
query social_network {
  city {
    name
    resident {
      message {
        id, day, content
      }
    }
  }
}

```

```

{ "cities": [ {
  "name": "Mol",
  "residents": [
    { "name": "Carl" },
    { "name": "Eve",
      "messages": [
        { "id": "M3", "day": "Sun", "content": "alright" },
        { "id": "M4", "day": "Tue", "content": "Hello" }
      ]
    }
  ]
} ] ] }

```



GraphQL

OrientDB & co.

2010

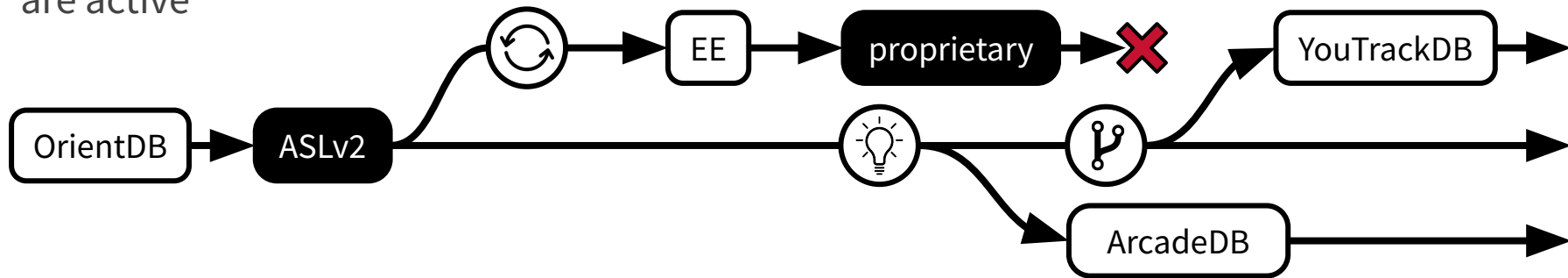
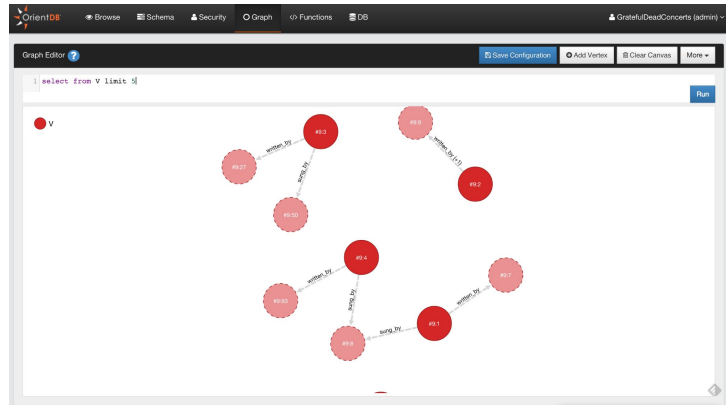
SQL dialect, Gremlin

User-friendly SQL dialect:

```
SELECT * FROM message m  
WHERE m.author.city.country = 'BE';
```

SAP acquired OrientDB, then abandoned it

Its open-source repository and two of its forks
are active



Microsoft Cosmos DB

2017

SQL dialect, Gremlin, etc.

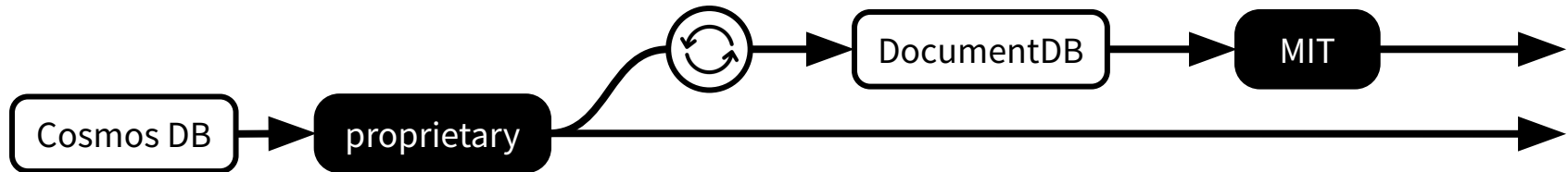
Distributed, schemaless NoSQL graph database

Powers ChatGPT in Azure

Also offers a SQL-like language, Gremlin, and a MongoDB-compatible API

NoSQL

```
SELECT VALUE {  
  "name": CONCAT(e.name.last, " ", e.name.first),  
  "department": e.department.name,  
  "emailAddresses": [  
    e.email  
  ]  
}  
FROM  
  employees e  
WHERE  
  STRINGEQUALS(e.department.name, "logistics", true)
```



 **Aerospike**

Gremlin

 **ArangoDB**

AQL

 **Dgraph**

DQL (\approx GraphQL)

LinkedIn Liquid

Datalog

 **SurrealDB**

SQL, GraphQL

 **TerminusDB**

Datalog, WOQL
(\approx GraphQL)

Hints for spotting graph serving systems

A graph database is likely a **graph serving** system if it is:

- backed by a key-value / document store
- called a “real-time graph database”
- categorized under “multi-model” graph database

Rank			DBMS	Database Model	Score	
Jan 2025	Dec 2024	Jan 2024			Jan 2025	Dec 2024
1.	1.	1.	Neo4j	Graph	43.69	+0.62
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model	Document store Graph DBMS Key-value store Wide column store Spatial DBMS	
3.	3.	3.	Aerospike	Multi-model		
4.	4.	4.	Virtuoso	Multi-model		
5.	5.	5.	ArangoDB	Multi-model		
6.	6.	6.	OrientDB	Multi-model		
7.	7.	10.	GraphDB	Multi-model		2.72
8.	8.	7.	Memgraph	Graph	2.62	-0.08

transactional

analytical



“MongoDB”

CAT II

CAT III

repeated joins =
n + 1 query problem

transactional

analytical



“MongoDB”

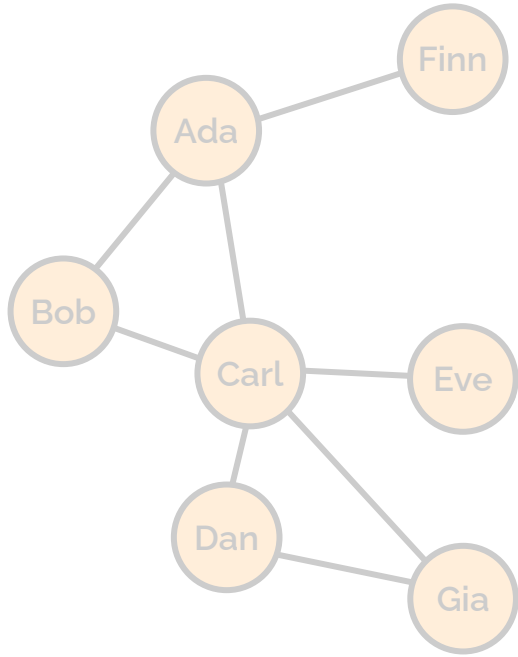
repeated joins =
n + 1 query problem

**classic graph
database**

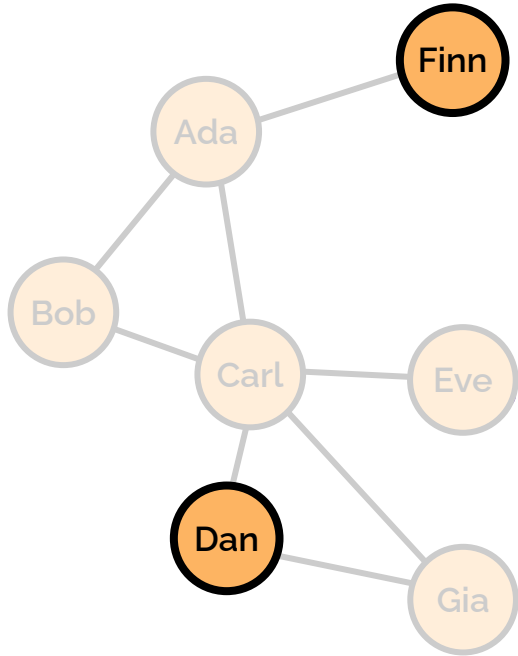
recursive joins =
path queries

CAT III

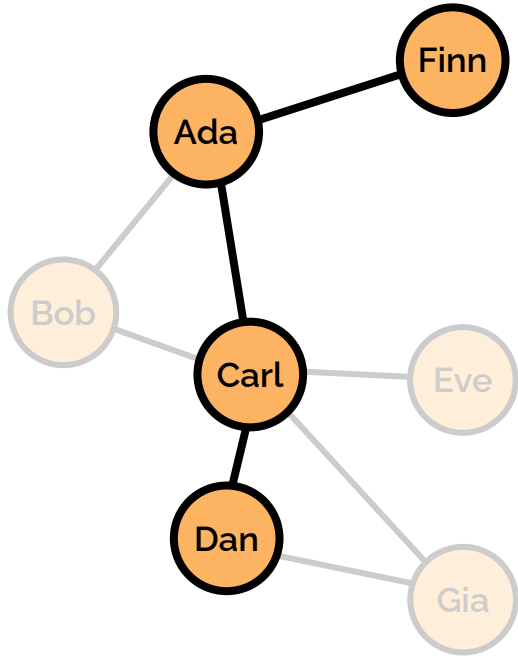
Shortest path finding



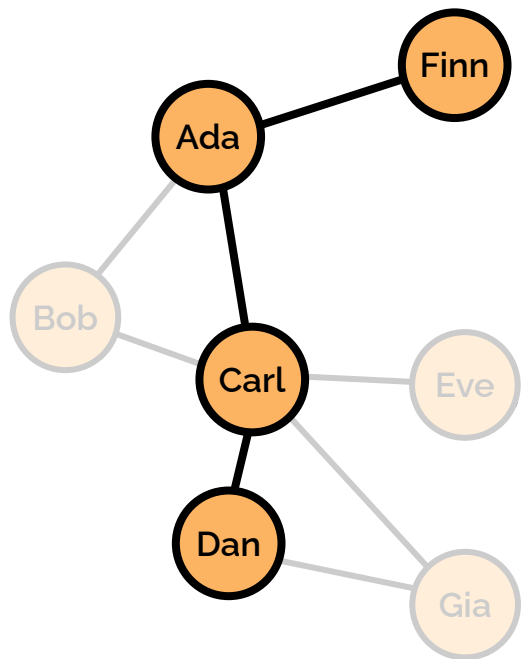
Shortest path finding



Shortest path finding

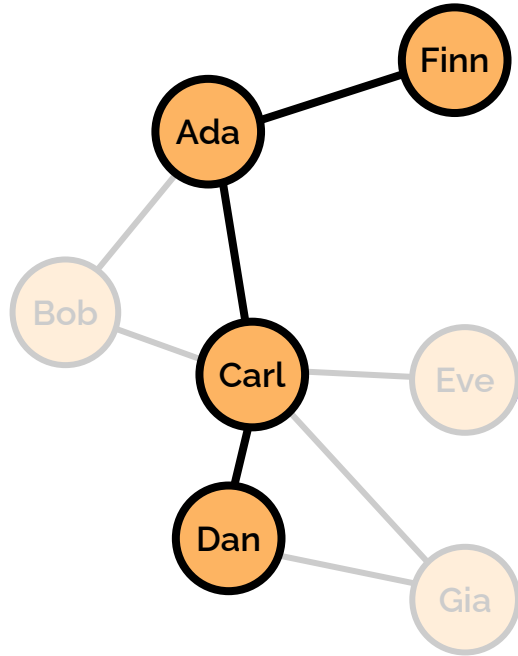


Shortest path finding in SQL:1999



```
WITH RECURSIVE paths(endNode, path, ended) AS (  
  SELECT p2 AS endNode, [p1, p2] AS path,  
         (p2 = 'Dan') AS ended FROM knows WHERE p1 = 'Finn'  
  UNION ALL  
  SELECT p2 AS endNode, array_append(path, p2) AS path,  
         max(p2 = 'Dan') OVER (ROWS BETWEEN UNBOUNDED PRECEDING  
                               AND UNBOUNDED FOLLOWING) AS ended  
  FROM paths JOIN knows ON paths.endNode = p1  
  WHERE NOT EXISTS (SELECT true FROM paths previous_paths  
                   WHERE list_contains(previous_paths.path, p2))  
                   AND NOT paths.ended)  
SELECT path FROM paths WHERE endNode = 'Dan';  
  
-- Result: [Finn, Ada, Carl, Dan]
```

Shortest path finding in Cypher



```
MATCH
  p = shortestPath(
    (:Person { name: 'Finn' })-[:knows*]-
    (:Person { name: 'Dan' })
  )
RETURN p;
```

Neo4j

2007

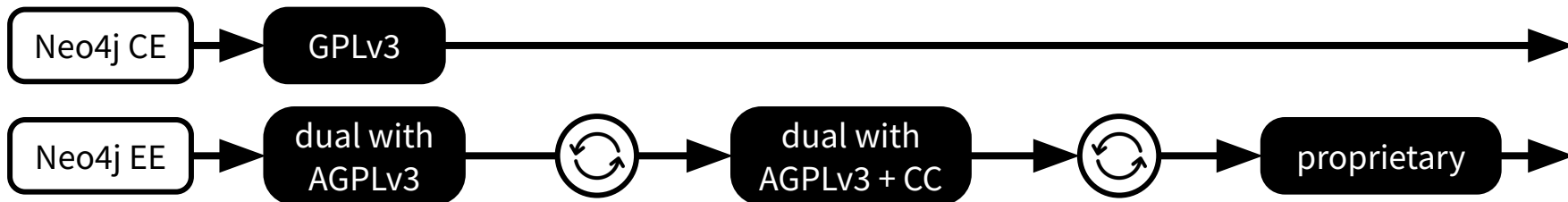
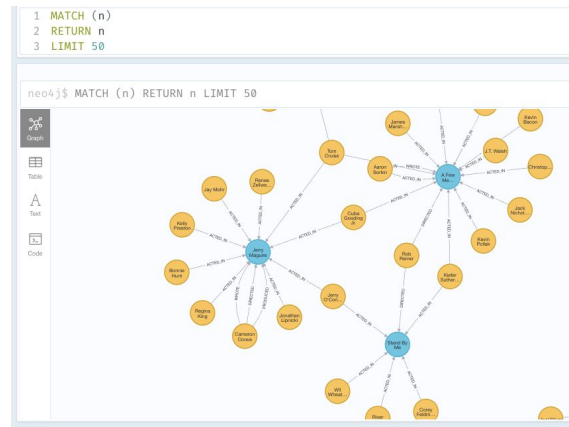
Cypher

The first modern graph database

- Cypher query language → openCypher → GQL
- Advocacy work: books, meetups, FOSDEM devrooms

Leans into analytics

- Analytics suite
- Parallel runtime in EE



Titan / JanusGraph / HugeGraph

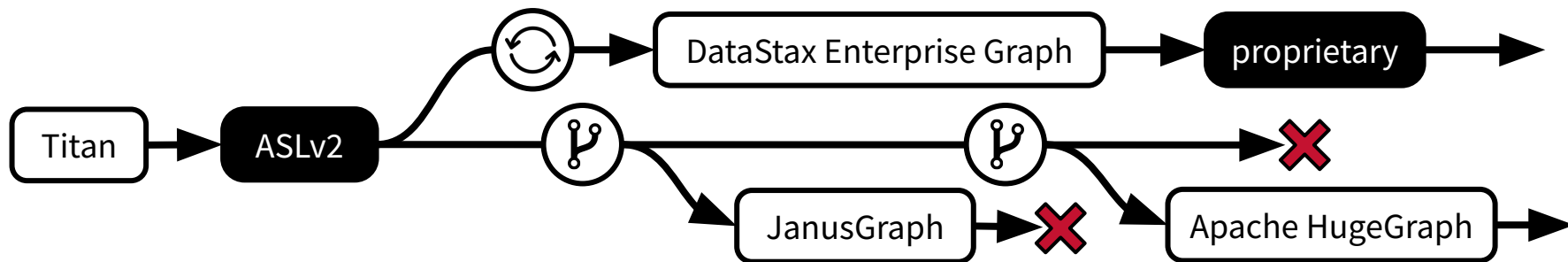
2012

Gremlin

Distributed graph database

Storage layer: RocksDB, Cassandra, MySQL / PostgreSQL, etc.

Workload: highly transactional setups





Amazon Neptune



GALAXYBASE



GraphScope

Cypher, Gremlin,
SPARQL

Cypher, Gremlin

Gremlin



MEM GRAPH



Nebulagraph



Google Cloud Spanner Graph



ULTIPA

Cypher

Cypher,
nGQL, GQL

SQL/PGQ,
GQL

UQL,
GQL



Cypher, Gremlin,
SPARQL



Cypher, Gremlin



Gremlin



Cypher



Cypher,
nGQL, GQL



SQL/PGQ,
GQL



UQL,
GQL

Extensions for RDBMS

- IBM Db2 Graph: Gremlin transpiled to SQL
- PostgreSQL AGE extension: openCypher transpiled to SQL
- PostgreSQL: SQL/PGQ (WIP)
- Oracle Database 23ai: SQL/PGQ
- Microsoft SQL Server Graph: SQL/PGQ-like custom syntax
- SAP HANA Graph: openCypher for queries, GraphScript for algorithms

transactional

analytical



“MongoDB”

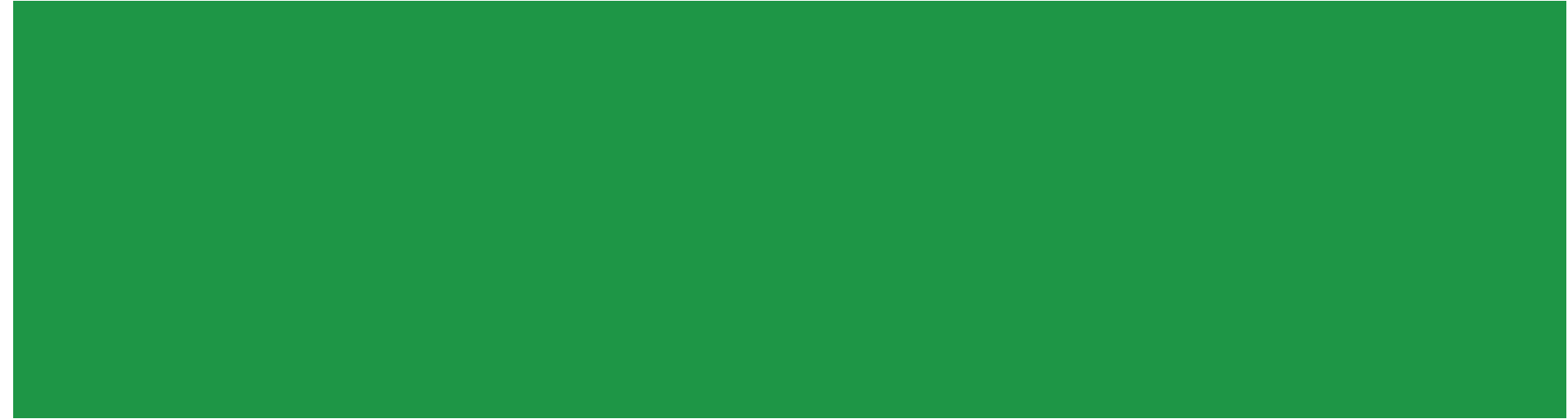
repeated joins =
n + 1 query problem

“Postgres”

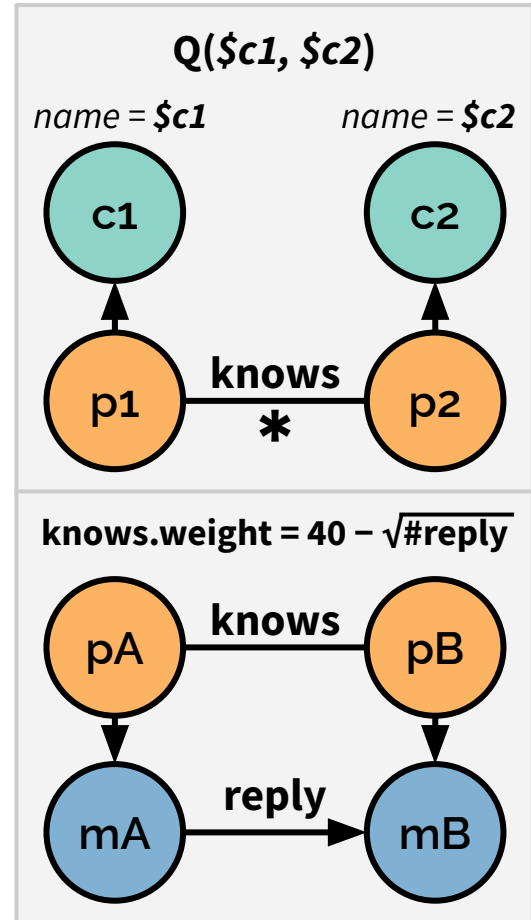
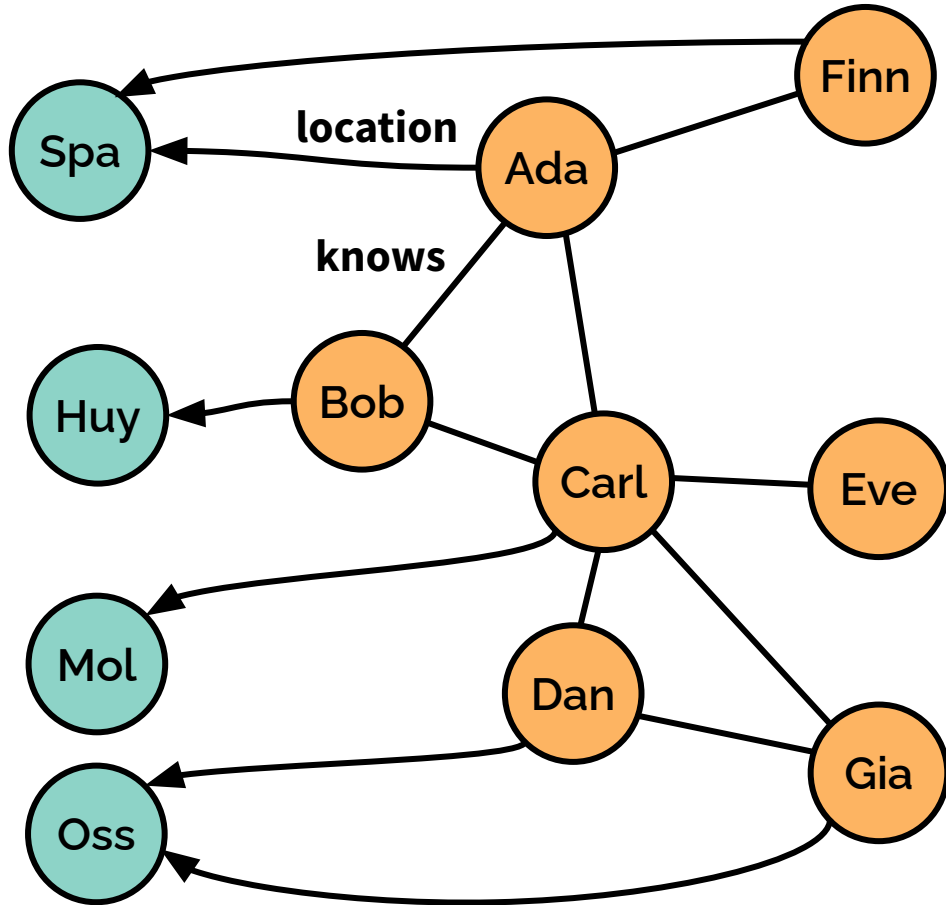
recursive joins =
path queries

CAT III

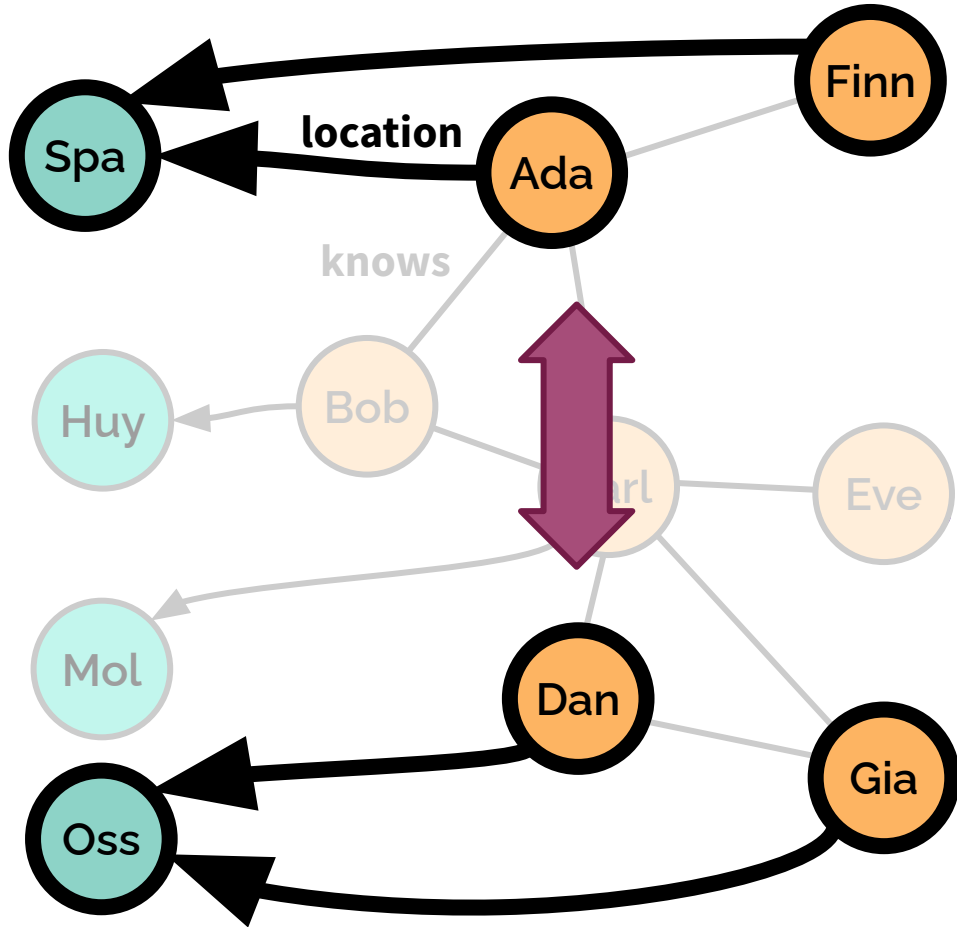
Graph BI 1: Cheapest path queries



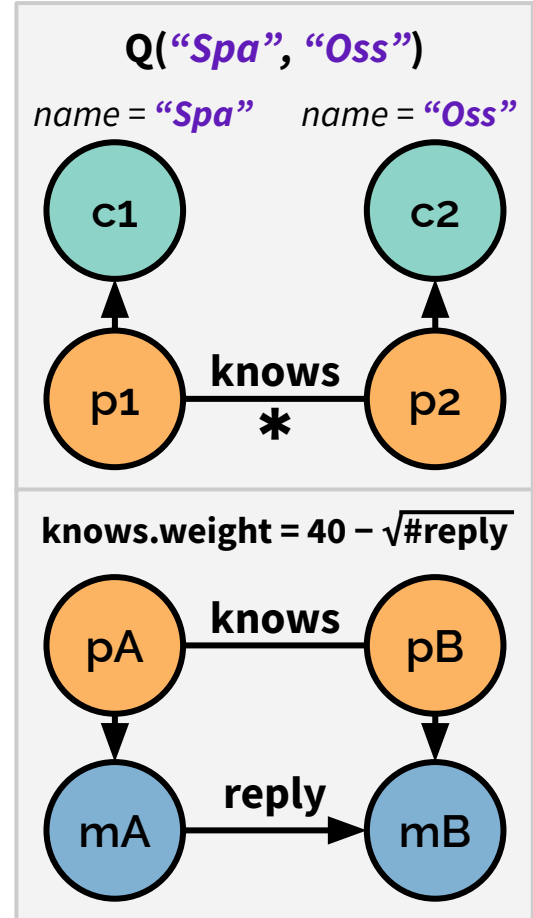
Find cheapest paths (weighted shortest)



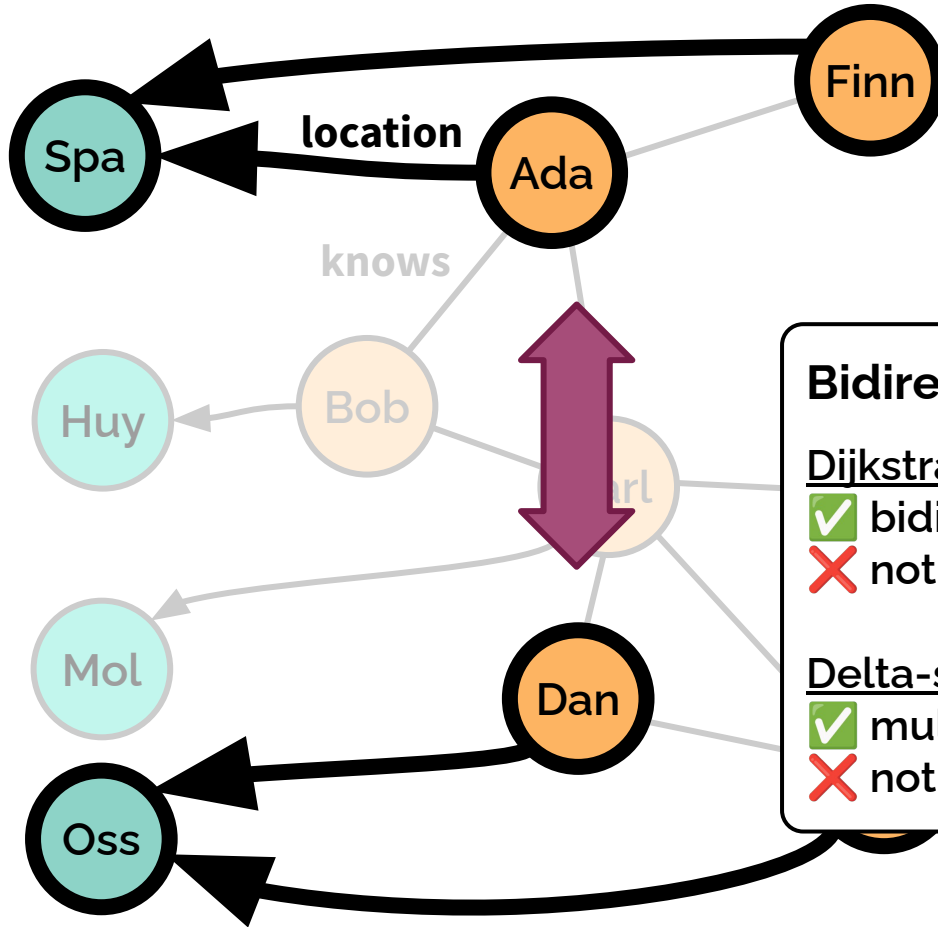
Find cheapest paths (weighted shortest)



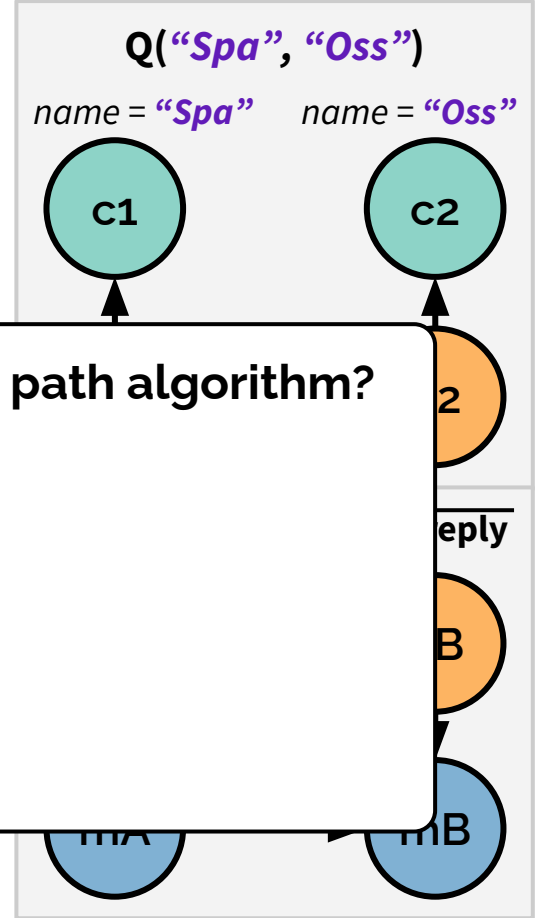
src	trg	w
Ada	Dan	...
Ada	Gia	...
Finn	Dan	...
Finn	Gia	...



Find cheapest paths (weighted shortest)



src	trg	w
Ada	Dan	...
Ada	Gia	...
Finn	Dan	...
Finn	Gia	...

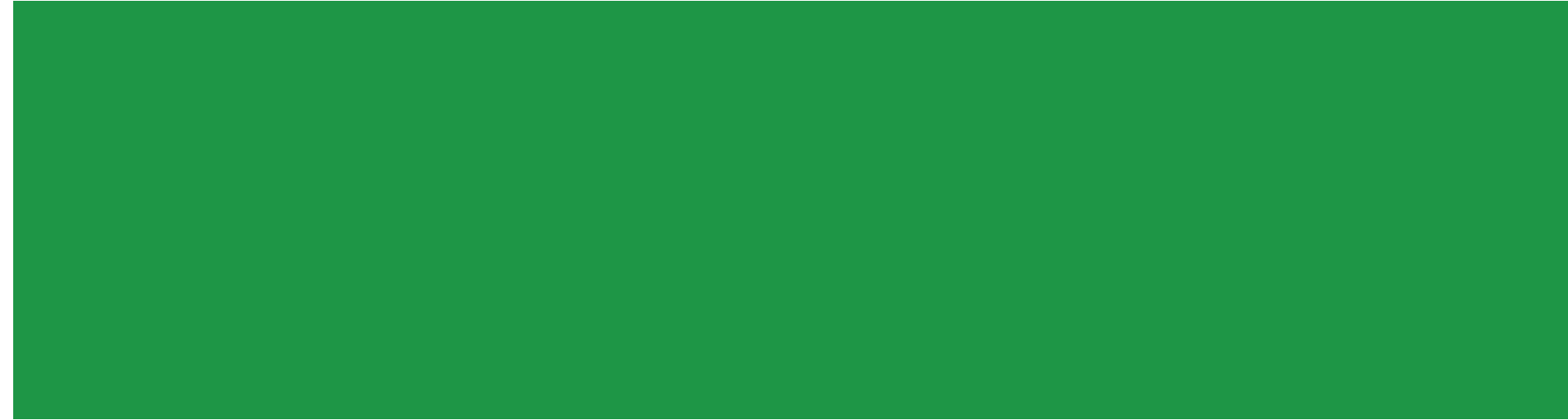


Bidirectional cheapest path algorithm?

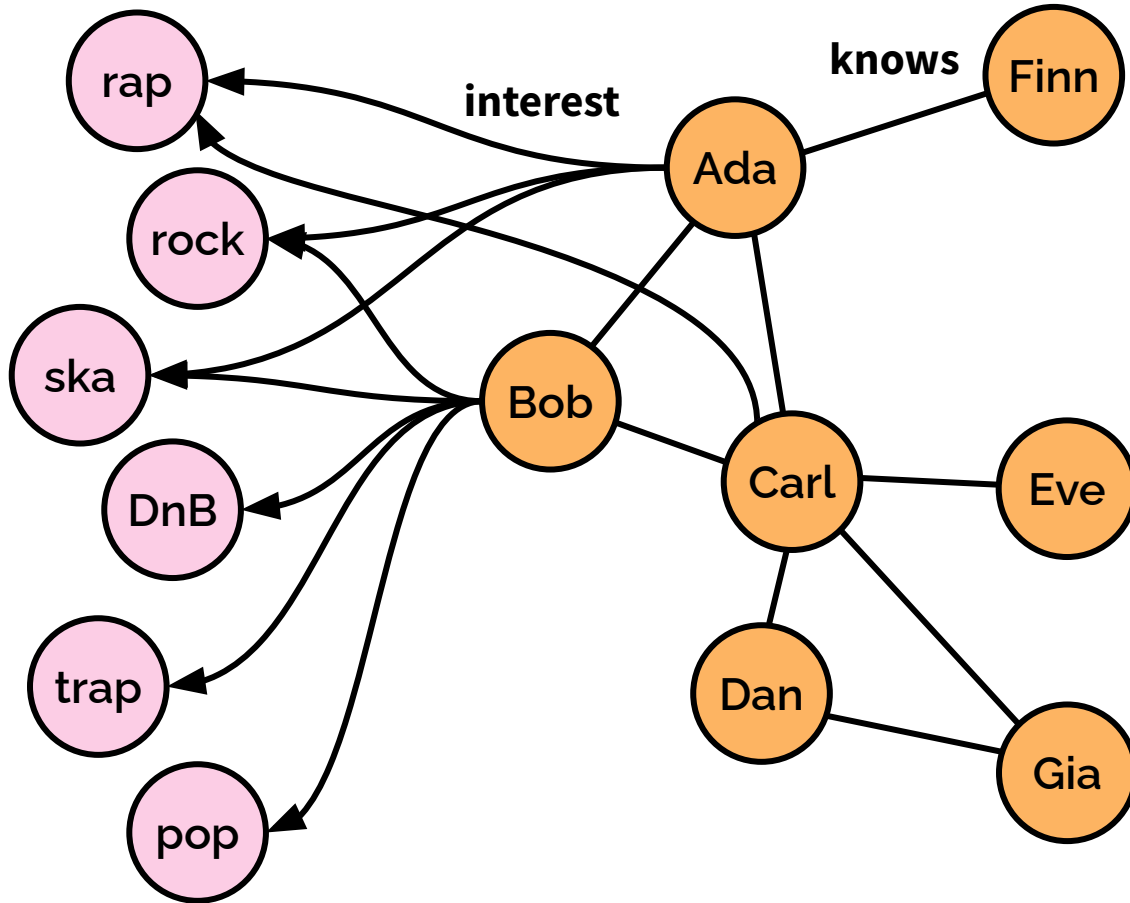
Dijkstra's algorithm
 ✓ bidirectional
 ✗ not multi-threaded

Delta-stepping
 ✓ multi-threaded
 ✗ not bidirectional

Graph BI 2: Cyclic queries

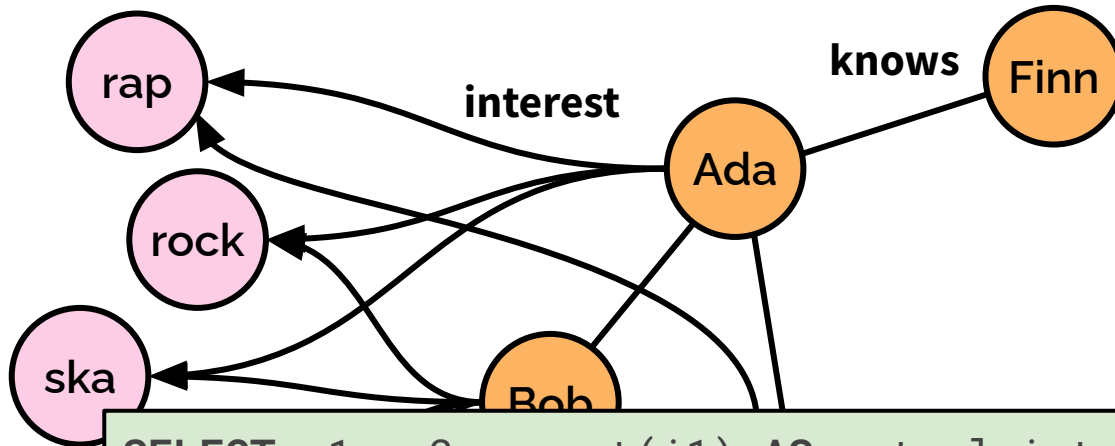


Cyclic query: Return shared interests



t1	p1	p2	t2
rap	Ada	Bob	rock
rock	Ada	Bob	rock
ska	Ada	Bob	rock
rap	Ada	Bob	ska
rock	Ada	Bob	ska
ska	Ada	Bob	ska
rap	Ada	Bob	DnB
rock	Ada	Bob	DnB
ska	Ada	Bob	DnB
rap	Ada	Bob	trap
rock	Ada	Bob	trap
ska	Ada	Bob	trap
rap	Ada	Carl	rap
rock	Ada	Carl	rap
ska	Ada	Carl	rap

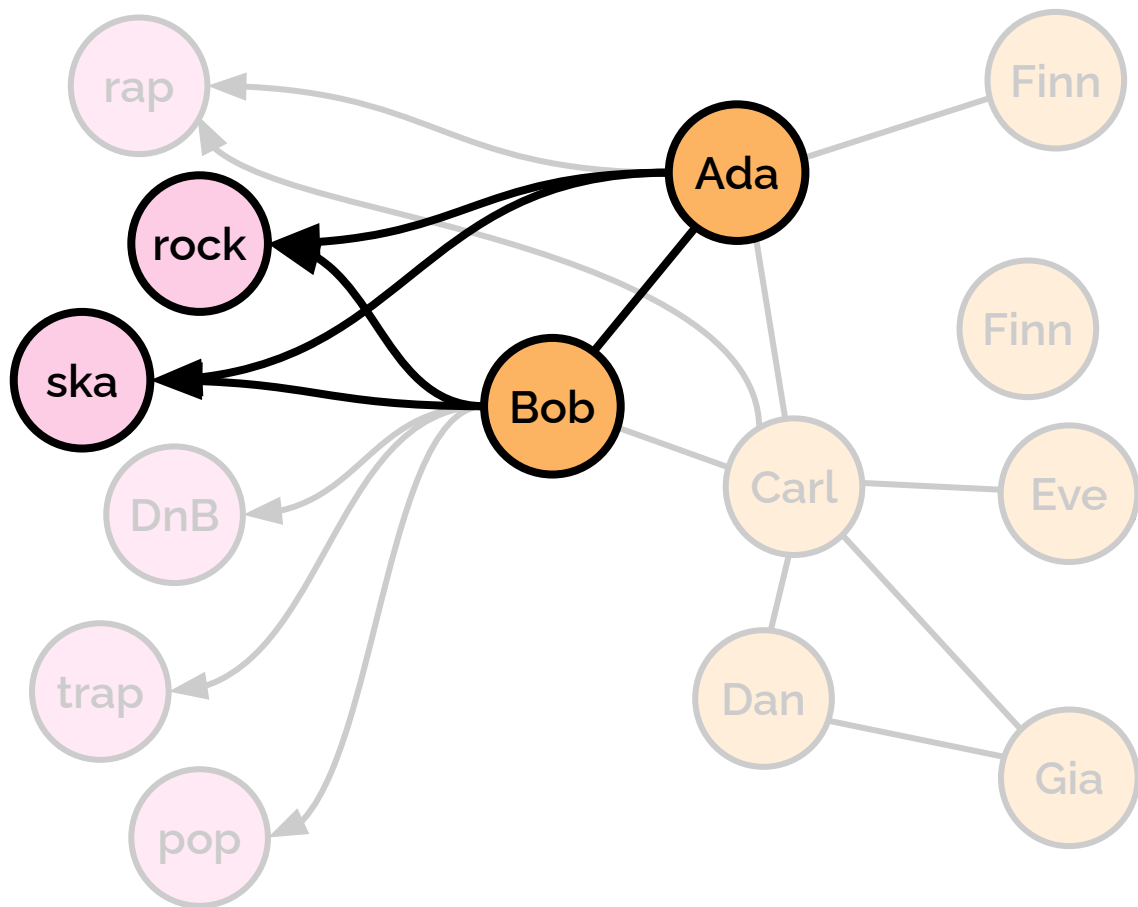
Cyclic query: Return shared interests



```
SELECT p1, p2, count(i1) AS mutual_int
FROM knows_undir ku
JOIN interest i1
  ON i1.person = ku.p1
JOIN interest i2
  ON i2.person = ku.p2
  AND i1.topic = i2.topic
WHERE p1 < p2
GROUP BY p1, p2;
```

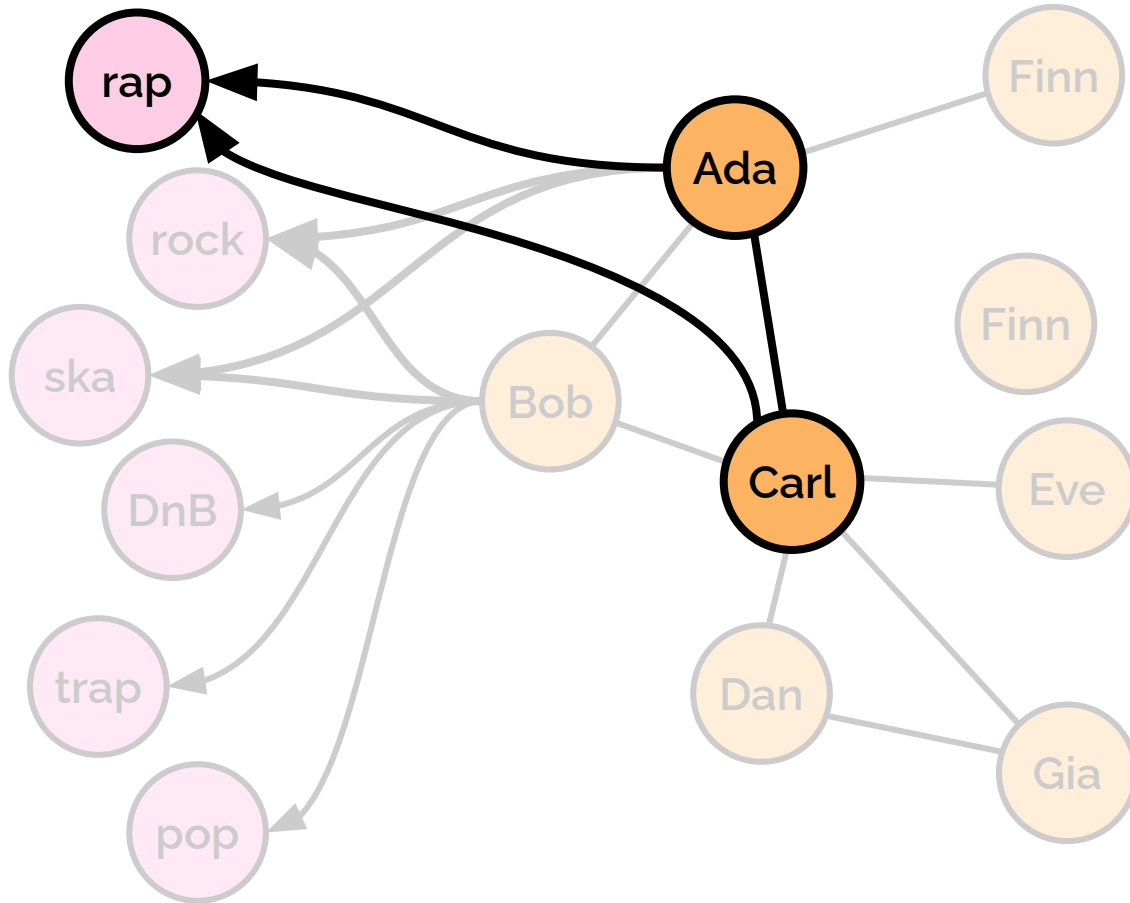
t1	p1	p2	t2
rap	Ada	Bob	rock
rock	Ada	Bob	rock
ska	Ada	Bob	rock
rap	Ada	Bob	ska
rock	Ada	Bob	ska
ska	Ada	Bob	ska
rap	Ada	Bob	DnB
rock	Ada	Bob	DnB
ska	Ada	Bob	DnB
rap	Ada	Bob	trap
rock	Ada	Bob	trap
ska	Ada	Bob	trap
rap	Ada	Carl	rap
rock	Ada	Carl	rap
ska	Ada	Carl	rap

Cyclic query: Return shared interests



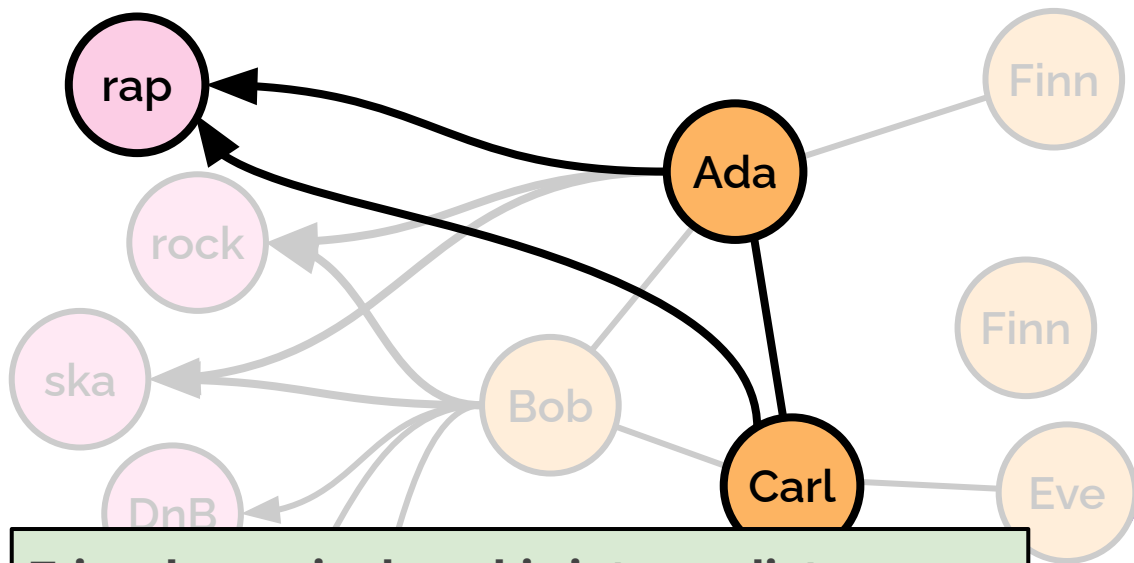
t1	p1	p2	t2
rap	Ada	Bob	rock
rock	Ada	Bob	rock
ska	Ada	Bob	rock
rap	Ada	Bob	ska
rock	Ada	Bob	ska
ska	Ada	Bob	ska
rap	Ada	Bob	DnB
rock	Ada	Bob	DnB
ska	Ada	Bob	DnB
rap	Ada	Bob	trap
rock	Ada	Bob	trap
ska	Ada	Bob	trap
rap	Ada	Carl	rap
rock	Ada	Carl	rap
ska	Ada	Carl	rap

Cyclic query: Return shared interests



t1	p1	p2	t2
rap	Ada	Bob	rock
rock	Ada	Bob	rock
ska	Ada	Bob	rock
rap	Ada	Bob	ska
rock	Ada	Bob	ska
ska	Ada	Bob	ska
rap	Ada	Bob	DnB
rock	Ada	Bob	DnB
ska	Ada	Bob	DnB
rap	Ada	Bob	trap
rock	Ada	Bob	trap
ska	Ada	Bob	trap
rap	Ada	Carl	rap
rock	Ada	Carl	rap
ska	Ada	Carl	rap

Cyclic query: Return shared interests



t1	p1	p2	t2
rap	Ada	Bob	rock
rock	Ada	Bob	rock
ska	Ada	Bob	rock
rap	Ada	Bob	ska
rock	Ada	Bob	ska
ska	Ada	Bob	ska
rap	Ada	Bob	DnB
rock	Ada	Bob	DnB
ska	Ada	Bob	DnB
rap	Ada	Bob	trap
rock	Ada	Bob	trap
ska	Ada	Bob	trap
rap	Ada	Carl	rap
rock	Ada	Carl	rap
ska	Ada	Carl	rap

Triangle queries have big intermediates

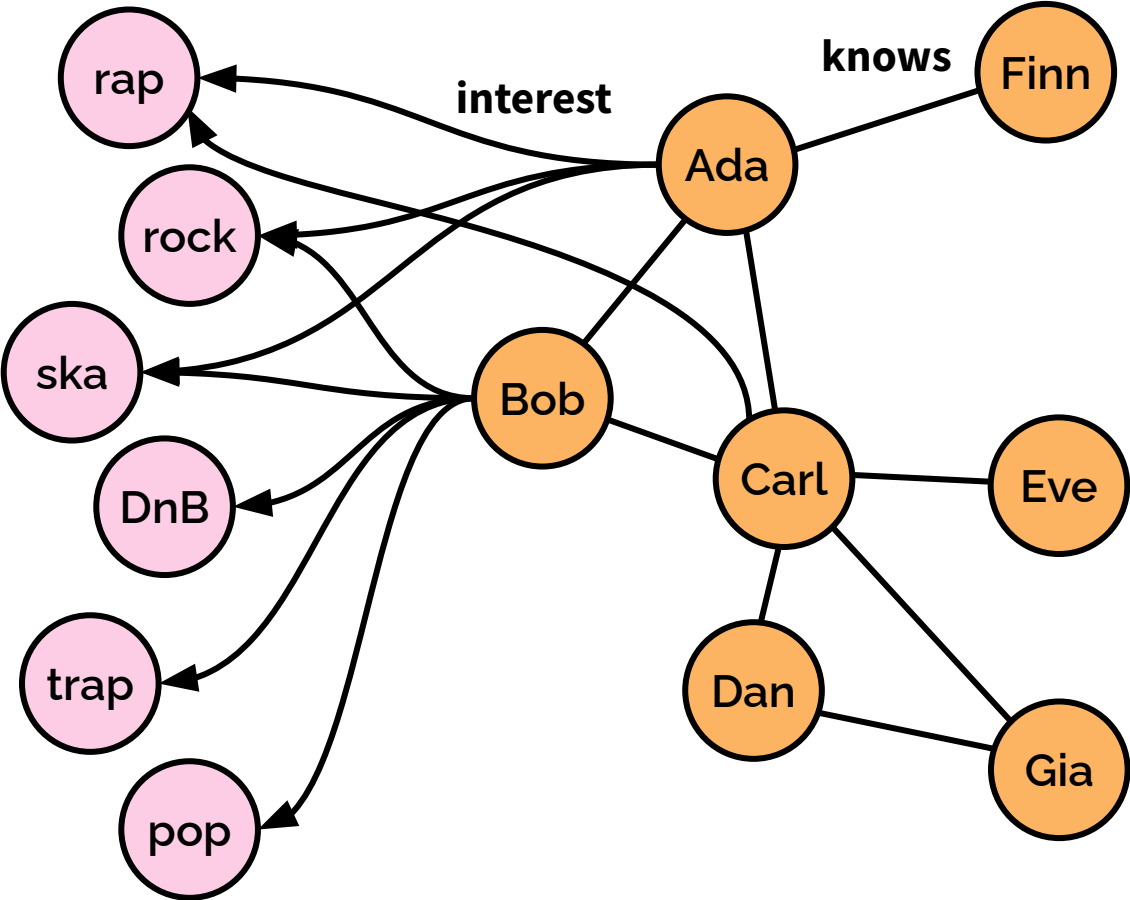
- interest1 \bowtie knows \bowtie interest2: $O(m^2)$
- \bowtie (interest1, knows, interest2): $O(m^{1.5})$

(Assuming $m = |\text{interest}| = |\text{knows}|$ here.)

Graph BI 3: Acyclic queries

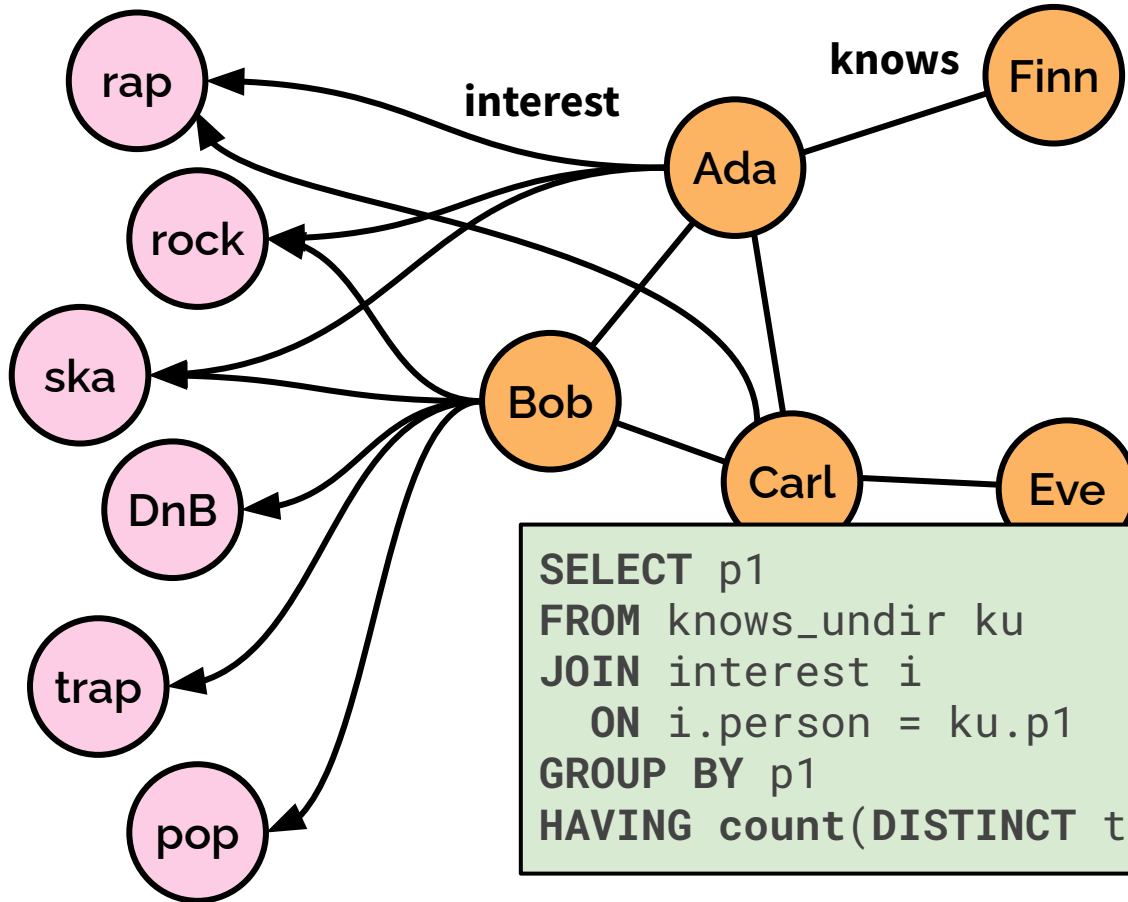


More interests than friends



Which person(s) have more topics of interests than friends?

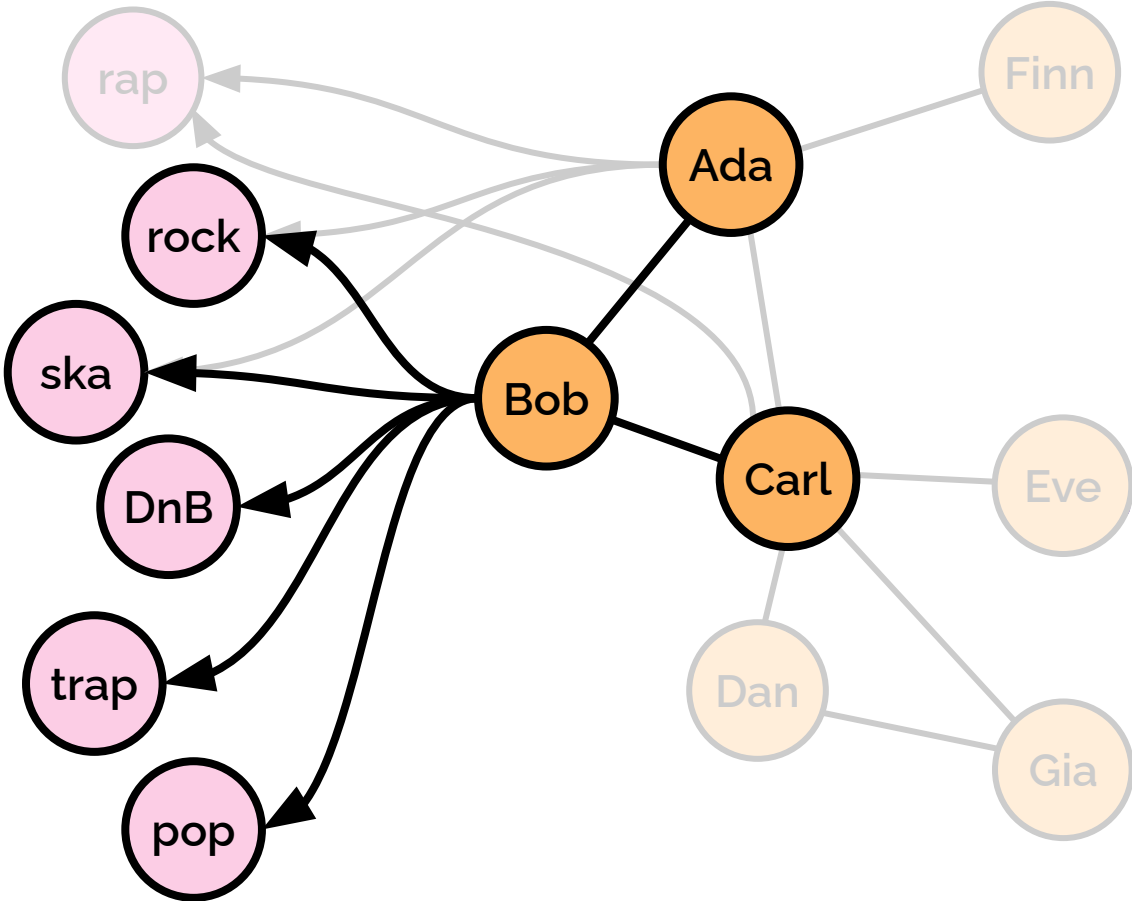
More interests than friends



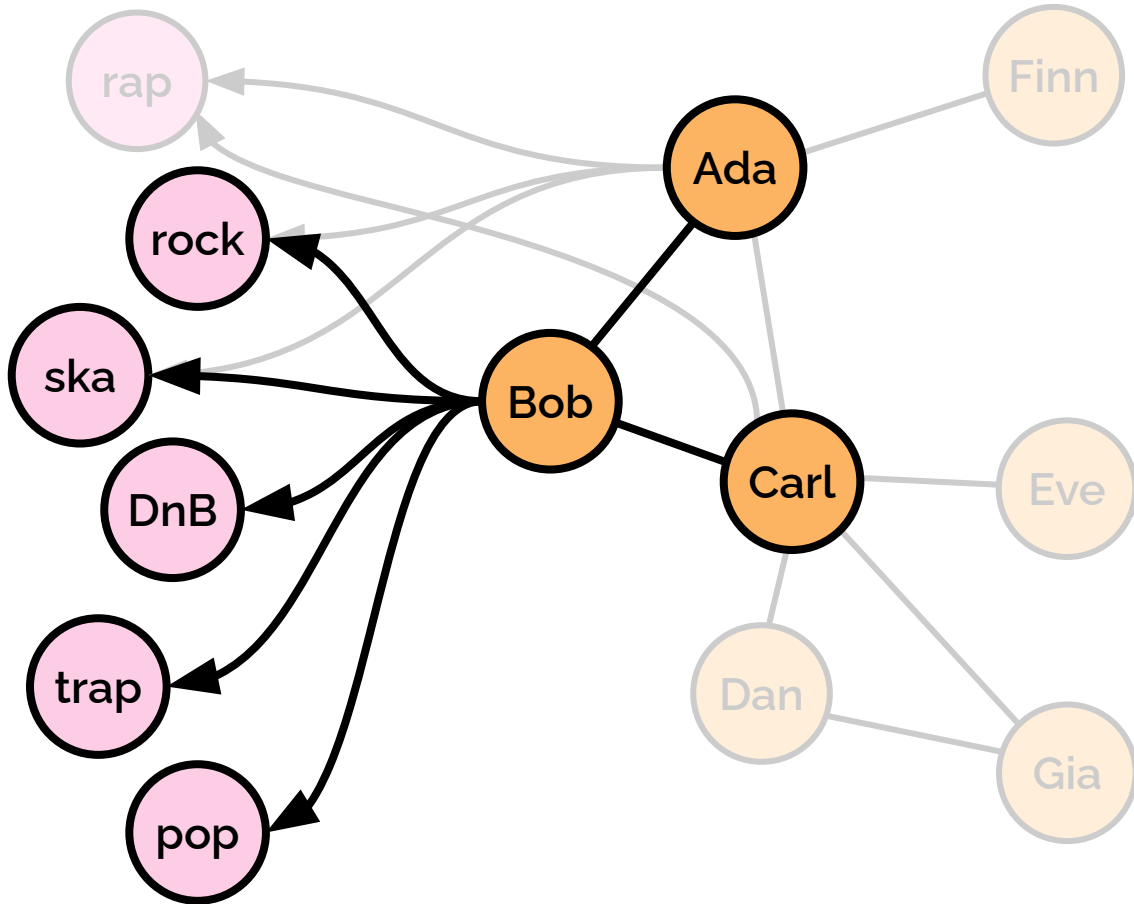
Which person(s) have more topics of interests than friends?

```
SELECT p1
FROM knows_undir ku
JOIN interest i
  ON i.person = ku.p1
GROUP BY p1
HAVING count(DISTINCT topic) > count(DISTINCT p2);
```

More interests than friends



More interests than friends



topic	p1	p2
rock	Bob	Ada
ska	Bob	Ada
DnB	Bob	Ada
trap	Bob	Ada
pop	Bob	Ada
rock	Bob	Carl
ska	Bob	Carl
DnB	Bob	Carl
trap	Bob	Carl
pop	Bob	Carl

Multi-valued dependency:

p1 \twoheadrightarrow **topic** and **p1** \twoheadrightarrow **p2**

Many-to-many joins introduce a predictable type of redundancy which we can compress away!

topic	p1	p2
rap	Ada	Bob
rock	Ada	Bob
ska	Ada	Bob
rap	Ada	Carl
rock	Ada	Carl
ska	Ada	Carl
rap	Ada	Finn
rock	Ada	Finn
ska	Ada	Finn
rock	Bob	Ada
ska	Bob	Ada
DnB	Bob	Ada
trap	Bob	Ada
pop	Bob	Ada
rock	Bob	Carl
ska	Bob	Carl
DnB	Bob	Carl
trap	Bob	Carl
pop	Bob	Carl
rap	Carl	Ada
rap	Carl	Bob

Factorization

Factorization is a lossless compression method.

flat to factorized



topic		p1		p2
{rap, rock, ska}	×	Ada	×	{Bob, Carl, Finn}
{rock, ska, DnB, trap, pop}	×	Bob	×	{Ada, Carl}
{rap}	×	Carl	×	{Ada, Bob}

```

SELECT p1
FROM knows_undir ku
JOIN interest i
  ON i.person = ku.p1
GROUP BY p1
HAVING count(DISTINCT topic) > count(DISTINCT p2);

```


topic	p1	p2
rap	Ada	Bob
rock	Ada	Bob
ska	Ada	Bob
rap	Ada	Carl
rock	Ada	Carl
ska	Ada	Carl
rap	Ada	Finn
rock	Ada	Finn
ska	Ada	Finn
rock	Bob	Ada
ska	Bob	Ada
DnB	Bob	Ada
trap	Bob	Ada
pop	Bob	Ada
rock	Bob	Carl
ska	Bob	Carl
DnB	Bob	Carl
trap	Bob	Carl
pop	Bob	Carl
rap	Carl	Ada
rap	Carl	Bob

Factorization

Factorization is a lossless compression method.

flat to factorized



topic		p1		p2
{rap, rock, ska}	×	Ada	×	{Bob, Carl, Finn}
{rock, ska, DnB, trap, pop}	×	Bob	×	{Ada, Carl}
{rap}	×	Carl	×	{Ada, Bob}

Workloads heavy on many-to-many joins could benefit from factorization but there are many open questions:

How to factorize long chains? Which queries benefit from factorization? How to implement it efficiently? How to return a factorized data structure to the client?

Where did we start from?

Advanced factorization methods (“d-representation”) also *returning compact graphs*

partOf ⌘ **City** ⌘ **location** ⌘ **Person** ⌘ **author** ⌘ **Message**

city name	post code	population	person name	birth year	message id	day	comment
Mol	2400	37,000	Carl	1986	NULL	NULL	NULL
Mol	2400	37,000	Eve	2001	M3	Sun	alright
Mol	2400	37,000	Eve	2001	M4	Tue	Hello

city name	post code	population		person name	birth year		message
Mol	2400	37,000	×	Carl	1986	×	NULL
				Eve	2001	×	{⟨M3, Sun, alright⟩, ⟨M4, Tue, Hello⟩}

Research papers

Worst-case optimal joins:

- 2013 **AGM bound**. SIAM J. Comput.
- 2014 **Worst-case optimal joins**. PODS
- 2019 **Vertex ordering in worst-case optimal joins**. VLDB
- 2020 **Hash-based worst-case optimal join implementations**. VLDB

Factorization:

- 2012 **FDB: A query engine for factorised relational databases**. VLDB
- 2015 **Size bounds for factorized representations of query results**. TODS
- 2024 **Optimizing queries with many-to-many joins**.
- 2025 **Adaptive factorization using linear-chained hash tables**. CIDR

Kùzu

2023

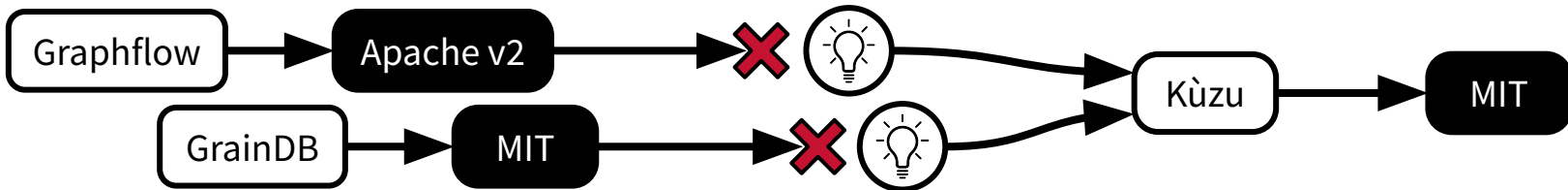
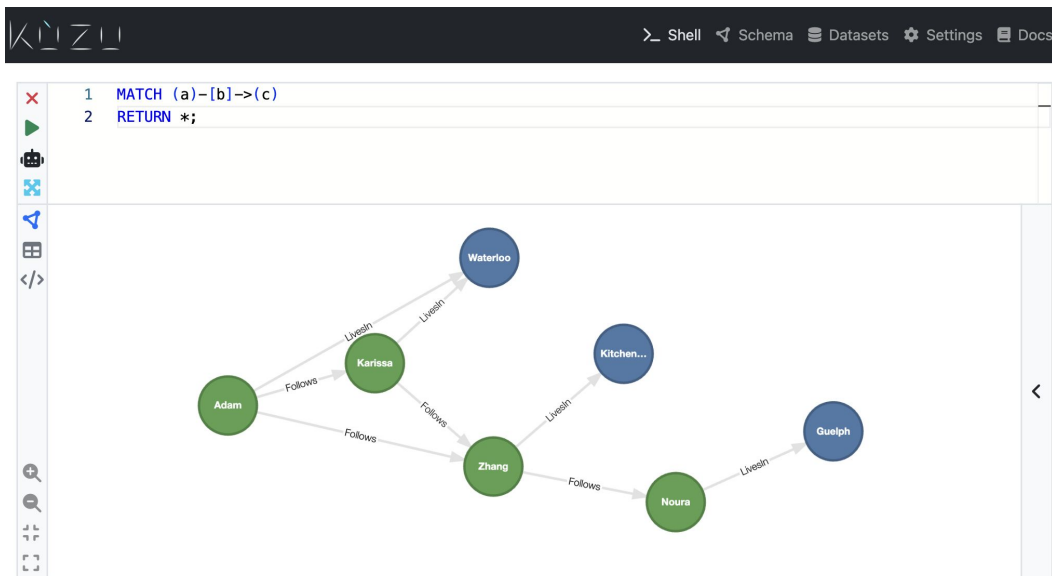
Cypher

Single-node, in-process system

Uses a relational backend

Supports Cypher

Strong focus on path queries, worst-case optimal joins, and factorization



TigerGraph

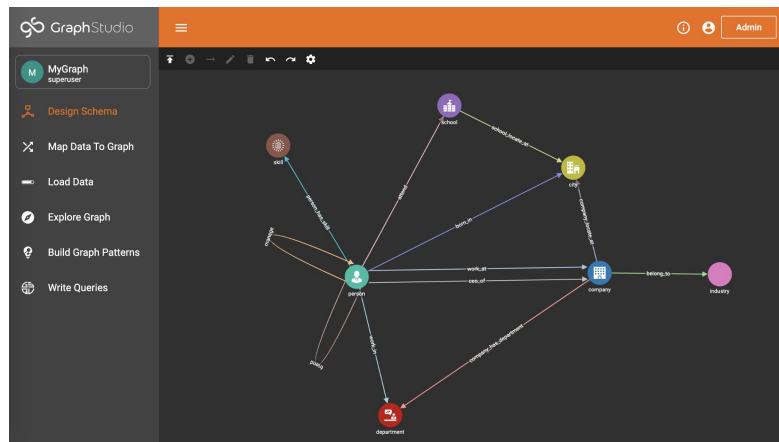
2012

GSQL

Focus on path queries

GSQL language:

```
CREATE QUERY hello(VERTEX<Person> p) {  
  Start = {p};  
  Result = SELECT dst  
    FROM Start:src -(knows:e)- Person:dst;  
  PRINT Result;  
}
```



TigerGraph

proprietary

transactional

analytical



“MongoDB”

repeated joins =
n + 1 query problem

“Postgres”

recursive joins =
path queries

“Teradata”

many-to-many joins =
complex recursive joins
cyclic graph patterns
long acyclic patterns

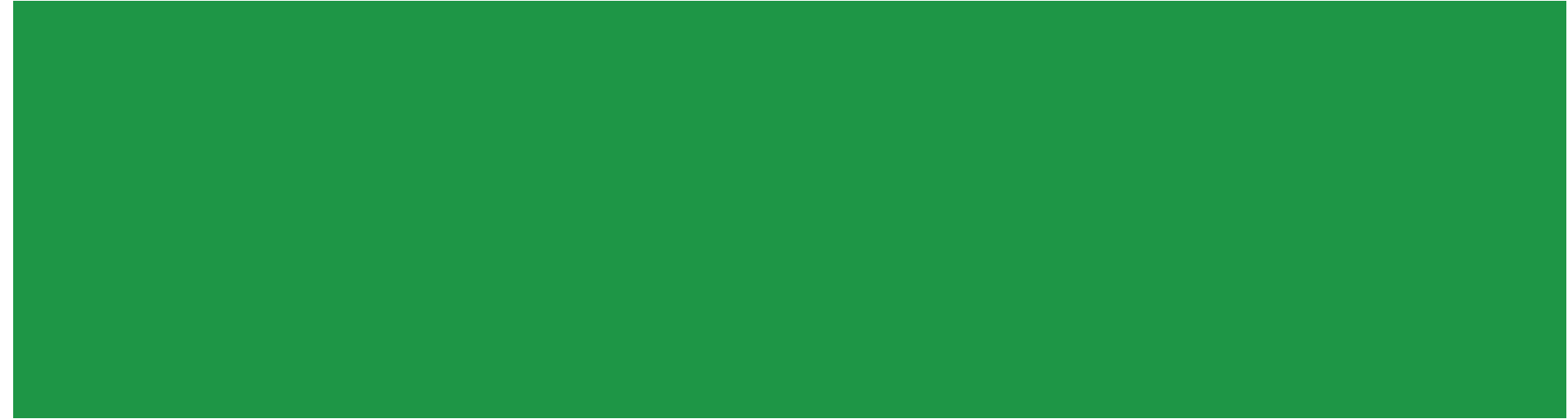
What about schema and distribution?



Schema? Single-node vs. distributed?

	single-node	distributed
no / optional schema	Neo4j CE OrientDB / ArcadeDB / YouTrackDB	Neo4j EE Cosmos DB
strict schema	Kùzu DuckPGQ	Titan / JanusGraph / HugeGraph TigerGraph

Benchmarks



LDBC: Linked Data Benchmark Council

Mission:

Accelerate progress
in graph data management

Membership:

~25 organizations
~100 individuals





HUAWEI CLOUD



Birkbeck
UNIVERSITY OF LONDON

*Sparsity



Predictable Labs, Inc.



Members



database companies



hardware vendors

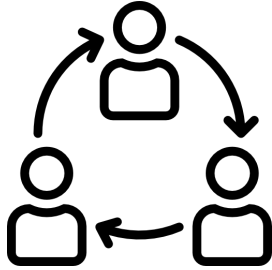


cloud providers

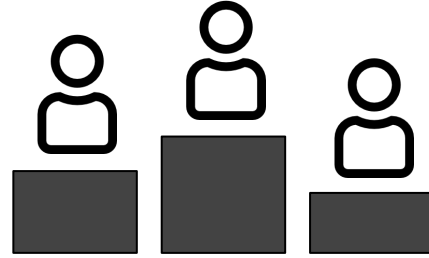


researchers

LDBC encourages members to...



collaborate
on standards



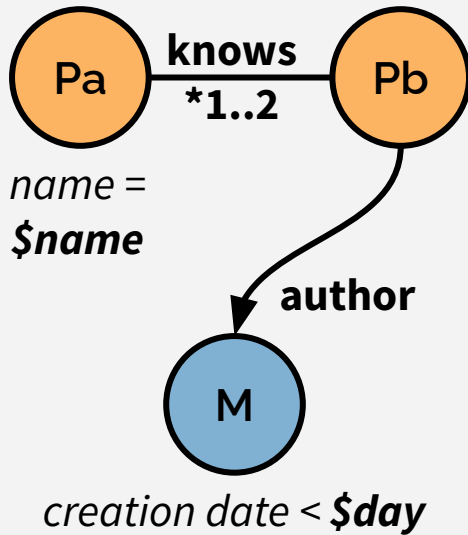
compete
on performance

Benchmarks and query languages



“Interactive” workload

Q9(\$name, \$day)



SQL:1992

```
SELECT DISTINCT m.id
FROM (
  SELECT k.p2id AS id
  FROM person Pa,
       knows k
  WHERE Pa.name = $name
       AND Pa.id = k.p1id
  UNION
  SELECT k2.p2id AS id
  FROM person Pa,
       knows k1,
       knows k2
  WHERE Pa.name = $name
       AND Pa.id = k1.p1id
       AND k1.p2id = k2.p1id
       AND k1.p1id <> k2.p2id
) Pb,
Message m
WHERE Pb.id = m.authorId
      AND m.creationDate < $day
```

SQL/PGQ (SQL:2023)

```
SELECT id
FROM GRAPH_TABLE (socialNetwork
  MATCH ANY ACYCLIC
  (Pa:Person WHERE Pa.name = $name)
  -[:knows]-{1,2} (Pb:Person)
  -[:author]-> (m:Message)
  WHERE m.creationDate < $day
  COLUMNS (m.id))
```

GQL

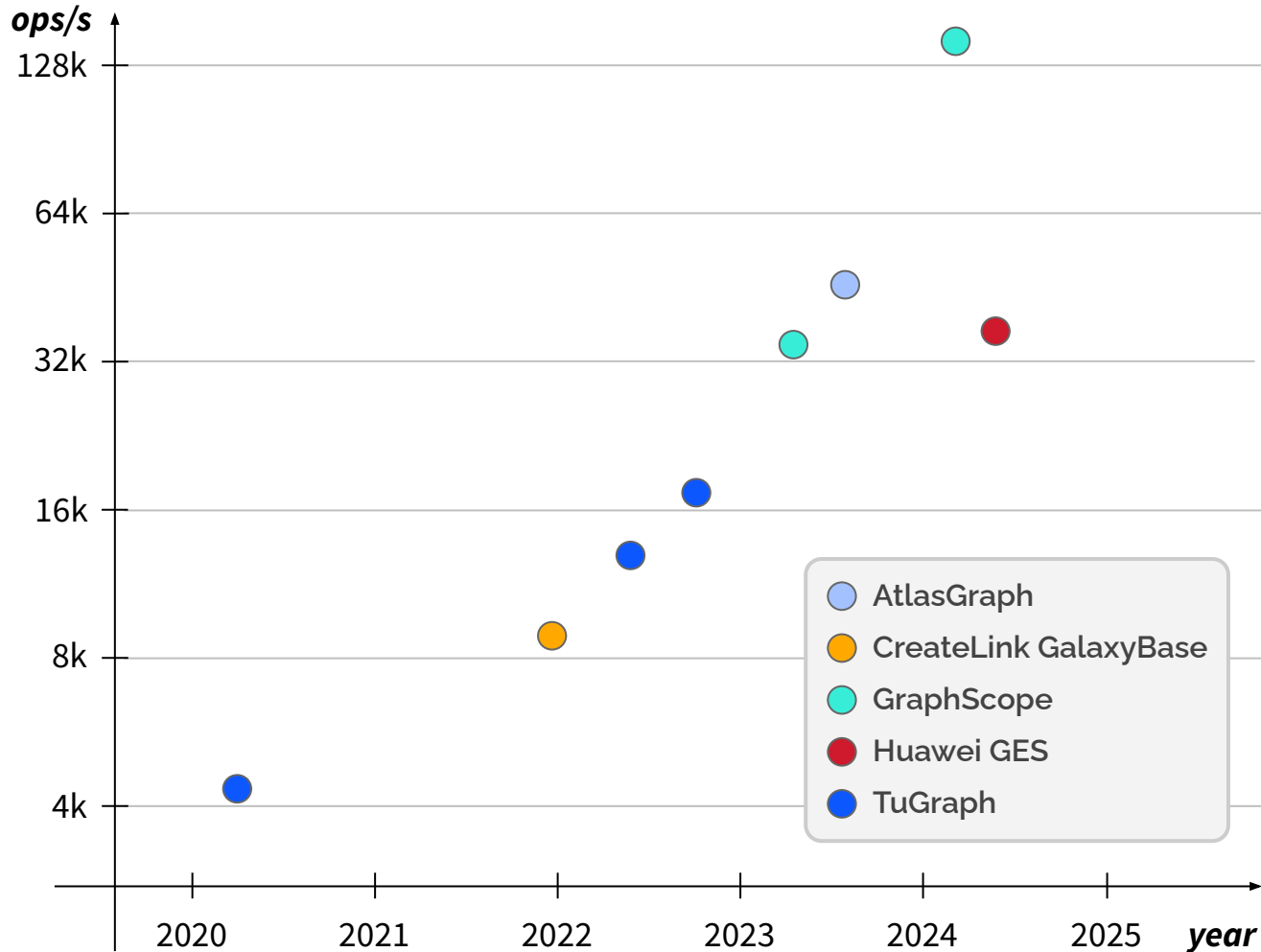
```
MATCH ANY ACYCLIC
  (Pa:Person WHERE Pa.name = $name)
  -[:knows]-{1,2} (Pb:Person)
  -[:author]-> (m:Message)
WHERE m.creationDate < $day
RETURN DISTINCT m.id
```

“Interactive” workload

SF100: 25× speedup in 4y

71× price-performance

All of these systems are developed by vendors based in China with a strong VC ecosystem

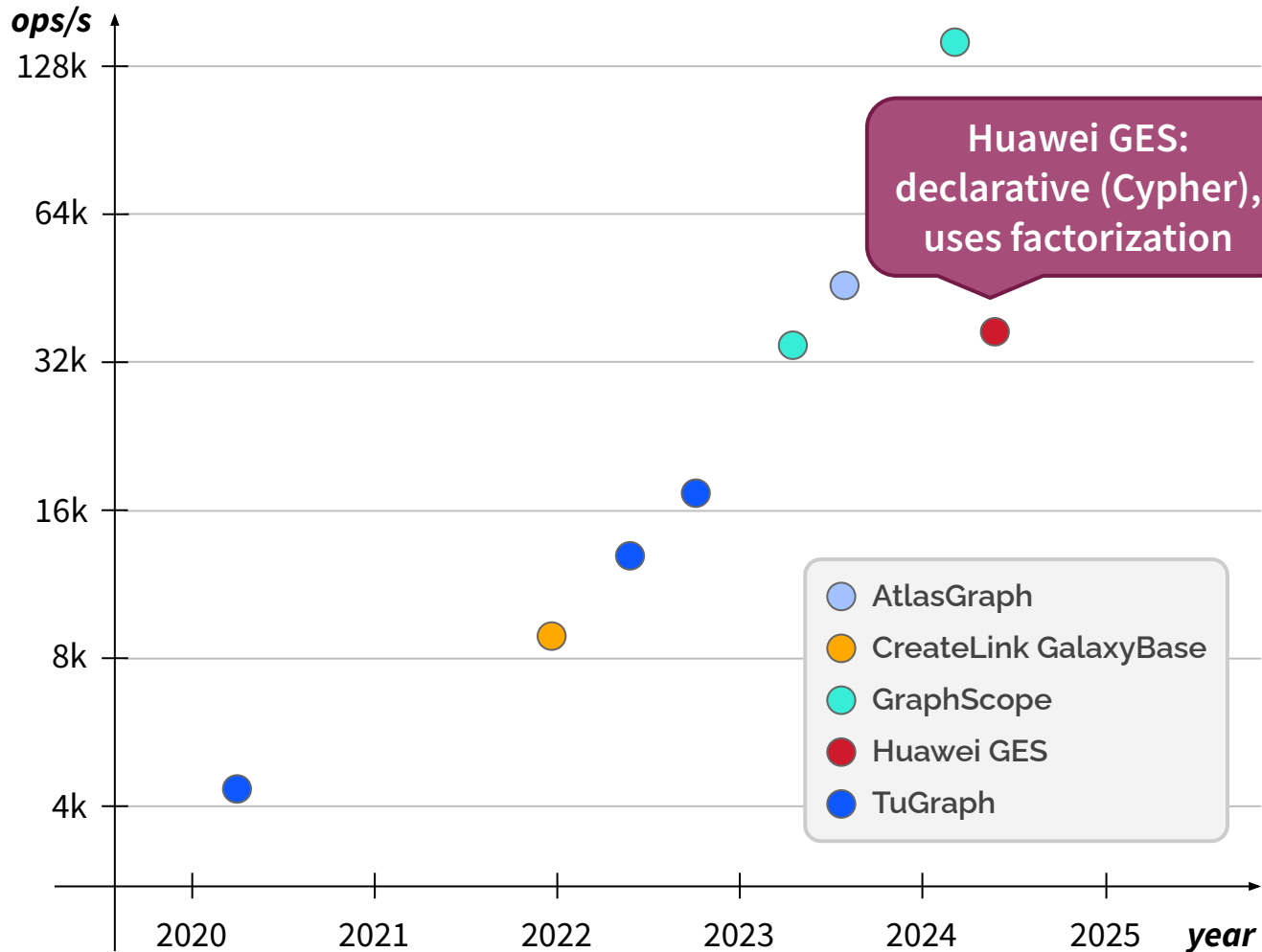


“Interactive” workload

SF100: 25× speedup in 4y

71× price-performance

All of these systems are developed by vendors based in China with a strong VC ecosystem



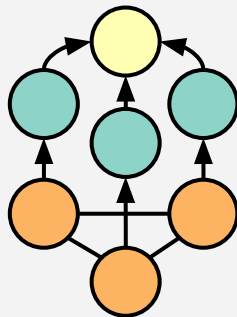
“Business Intelligence” workload

Analytical workload

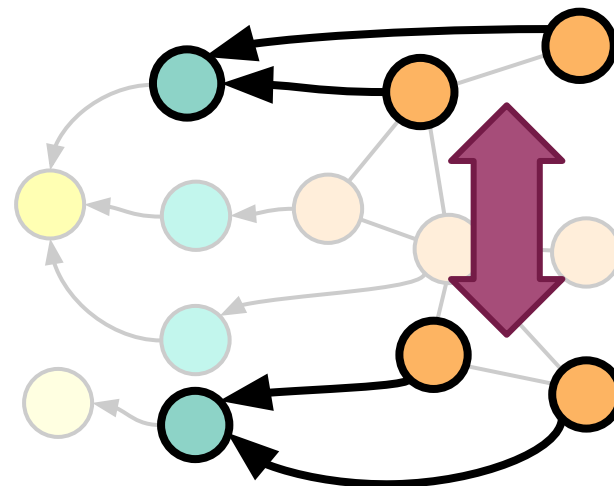
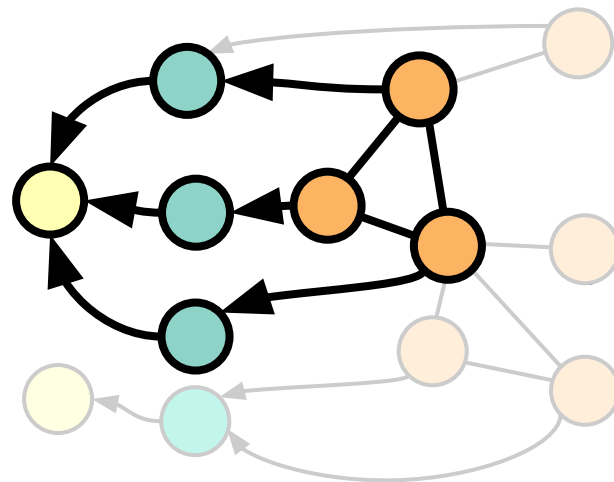
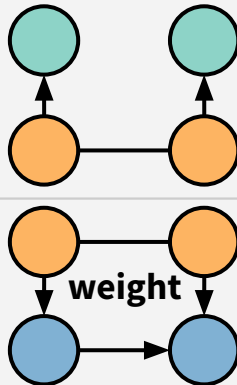
Metric 1: Power

Metric 2: Throughput

Q11(\$ctry)



Q19(\$c1, \$c2)



“Business Intelligence” workload

Analytical workload

Metric 1: Power

Metric 2: Throughput

Audited results

Scale factors



100

1,000 (×3)

10,000



30,000

More audits coming this year!

Financial Benchmark

Transactional workload

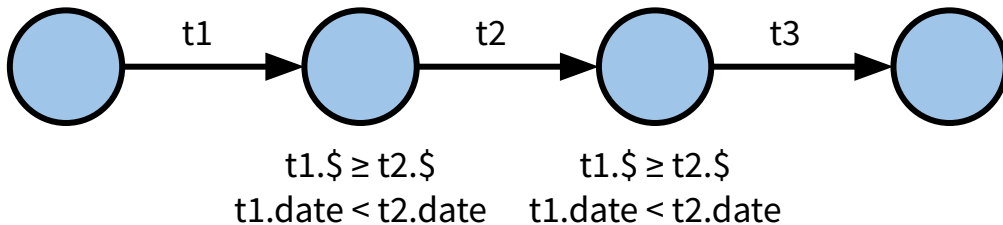
Metric: Throughput

Target: Distributed systems

Relaxed consistency requirements

Developed by the Ant Group, Create Link, Ultipa, etc.

- Strict latency requirements (P99 < 100 ms)
- Relaxed consistency guarantees
- Truncation (sampling) on most recent edges
- Interesting queries, e.g. REM path queries (Regular Expression with Memory)



Using the benchmarks



Benchmark kit

Specification

Academic paper


Data generator

Pre-generated data sets

Driver

2+ implementations


arXiv:2001.02299v8 [cs.DB] 9 Nov 2022



The LDBC Social Network Benchmark

(version 2.2.1)

The specification was built on the source code available at https://github.com/ldbc/ldbc_snb_docs/releases/tag/v2.2.1



The LDBC Social Network Benchmark

Business Intelligence Workload

Gábor Szárnyas CWI gszar@wi.kub.nl	Jack Waudby Newcastle University j.waudby@ncl.ac.uk	Benjamin A. Steer Pontery ben.steer@pontery.com	David Snakillas Independent contributor david.snakillas@gmail.com
Altan Birlir Technische Universität München altan.birlir@tum.de	Mingqi Wu TigerGraph mingqi.wu@tigergraph.com	Yuchen Zhang TigerGraph yuchen.zhang@tigergraph.com	Peter Boscé CWI boscep@wi.nl

ABSTRACT
The Social Network Benchmark Business Intelligence workload (SNB BI) is a comprehensive graph OLAP benchmark targeting analytical data systems capable of supporting graph workloads. This paper tracks the evolution of almost a decade of research in academic and industry via the Linked Data Benchmark Council (LDBC). SNB BI advances the state-of-the-art in specific, and available, analytical database benchmarks in many aspects. It has a sophisticated data generator, implemented on a scalable distributed infrastructure, that produces a social graph with small-world phenomena, whose value properties follow skewed and correlated distributions and whose values correlate with structure. This is a temporal graph where all nodes and edges follow lifespan-based rules with temporal skew enabling realistic and consistent temporal inserts and (reversible) deletes. The query workload exploiting this skew and correlation is based on LDBC's "choker point"-driven design methodology and will receive technical and scientific improvements in future (graph) database systems. SNB BI includes the first adoption of "parameter queries" in an analytical benchmark, a technique that ensures stable runtimes of query variants across different parameter values. Top performance metrics characterize peak single-query performance (power) and sustained execution query throughput. To demonstrate the portability of this benchmark, we present experimental results on a relational and a graph DBMS. Note that there do not constitute an official LDBC Benchmark Benchmark – only analysis results can use this trademarked term.

PNB LDBC Review Panel:
Gábor Szárnyas, Jack Waudby, Benjamin A. Steer, David Snakillas, Altan Birlir, Mingqi Wu, Yuchen Zhang, and Peter Boscé. The LDBC Social Network Benchmark Business Intelligence Workload (PNB BI) will be doi:10.14278/lldb.2022.02299v8

PNB LDBC Author Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/ldbc_snb_3rd_release/tag/v1.3

This work is licensed under the Creative Commons Attribution 4.0 International License. For more information, see <http://creativecommons.org/licenses/by/4.0/>. This work is derived from the LDBC Social Network Benchmark (LDBC SNB) project, which was sponsored by the LDBC Review Panel. Copyright is held by the respective Publications Offices licensed by the LDBC Review Panel.

Proceedings of the LDBC Symposium, Vol. 16, No. 1 (2022) 2022-2027.
doi:10.14278/lldb.2022.02299v8

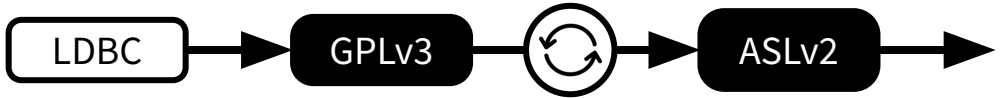
Table 1: The SNB BI workload fills in the space between LDBC SNB Interactive and LDBC Graphalytics. It is a graph OLAP workload focusing on queries on a labelled attributed graph with temporal changes (inserts and deletes), inserting systems with domain-specific query languages. We denote the data models and features covered, and whether a language is capable of implementing and allowed to implement a given benchmark. Notation: (C) yes, (N) no, (Q) limited coverage.

LDBC Workload	SNB BI		OLAP		Graph		Analytics	
	Yes	No	Yes	No	Yes	No	Yes	No
Labelled attributed graph	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Temporal operations	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Label semantics	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Ad-hoc query languages	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Domain-specific query languages	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Insert-only operations	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Query languages	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
OLAP workloads	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
SQL, SQL-PPG, Cypher, etc.	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
PNB BI, graph streams	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Temporal OLAP	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗

1 INTRODUCTION

Analyzing the connection patterns in graphs is a steadily expanding use case in data analysis and is growing in size and complexity in importance [17]. It is reflected in the increasing role of graph-based data as represented in data models such as (semi) RDF and increasingly property graphs [15]. While graph analytics is of the associated well-developed graph-oriented query engines that manage data representing social networks, telecommunication networks, and enterprise knowledge graphs [16], graph analytics are also found in traditional relational data warehouses and modern data lakes, where multi-graphs link to the continuous, partitioned, linked between tables that refer to each other through joins along relationships, say along many-to-many relationships. Practitioners, data system builders, and researchers are increasingly focusing on graph analytics questions [16], performing tasks such as fraud detection, recommendation, historical analysis, and real-time analytics.

The Linked Data Benchmark Council¹ To organize the evaluation of the modern graph data management tools, a group of industry and academic organizations founded the Linked Data Benchmark Council (LDBC) in 2012, originally as a European Union funded project.

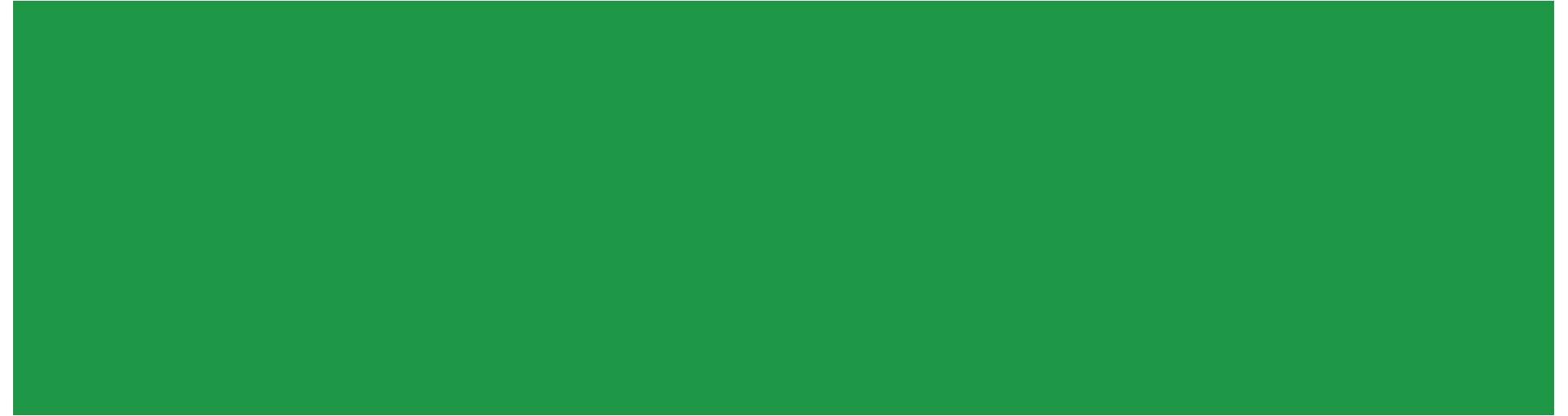


LDBC has “TPC-grade” benchmarks

- Auditing: an old model but it's still relevant
- Certified auditor, full disclosure report
- 3-year total cost of operations (licenses, support)
- Multi-week auditing process
- 5 years of auditing: published ≈ 50 results and had to retract 0

<ul style="list-style-type: none">○ System: GraphScope Flex 0.26.1○ Test sponsor: Alibaba Cloud○ Date: 2024-05-14	100	Alibaba Cloud ecs.r8a.16xlarge 64×AMD EPYC 9T24 @ 3.7GHz vCPUs, 512GiB RAM	130,098.36 ops/s
<ul style="list-style-type: none">○ Queries implemented in: C++○ System cost: 738,724 RMB (102,128.22 USD)	300	Alibaba Cloud ecs.r8a.16xlarge 64×AMD EPYC 9T24 @ 3.7GHz vCPUs, 512GiB RAM	131,263.87 ops/s

Challenges in the graph database space



Decline

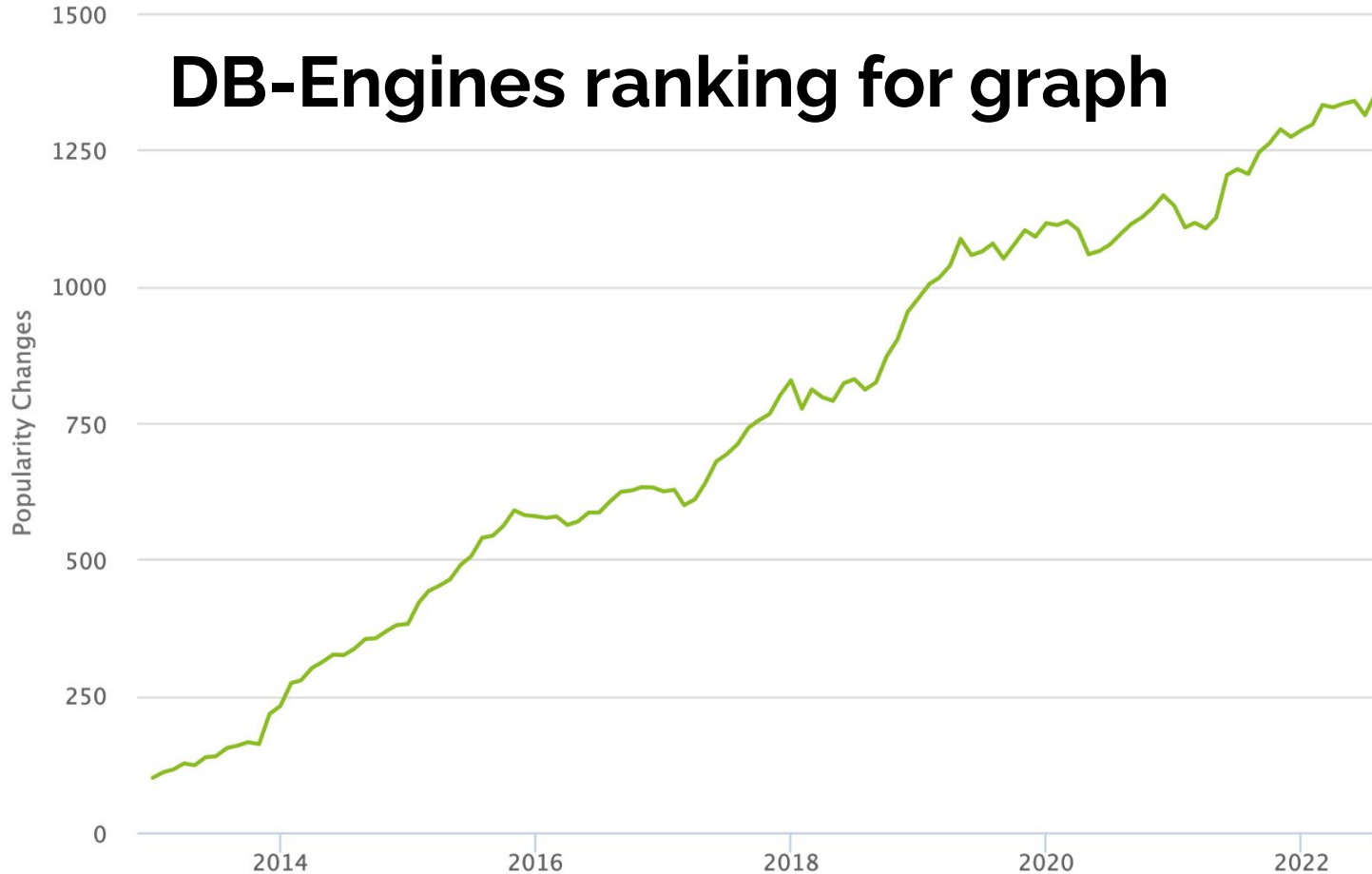
The hype cycle moved over to AI

The confusion doesn't help anyone!

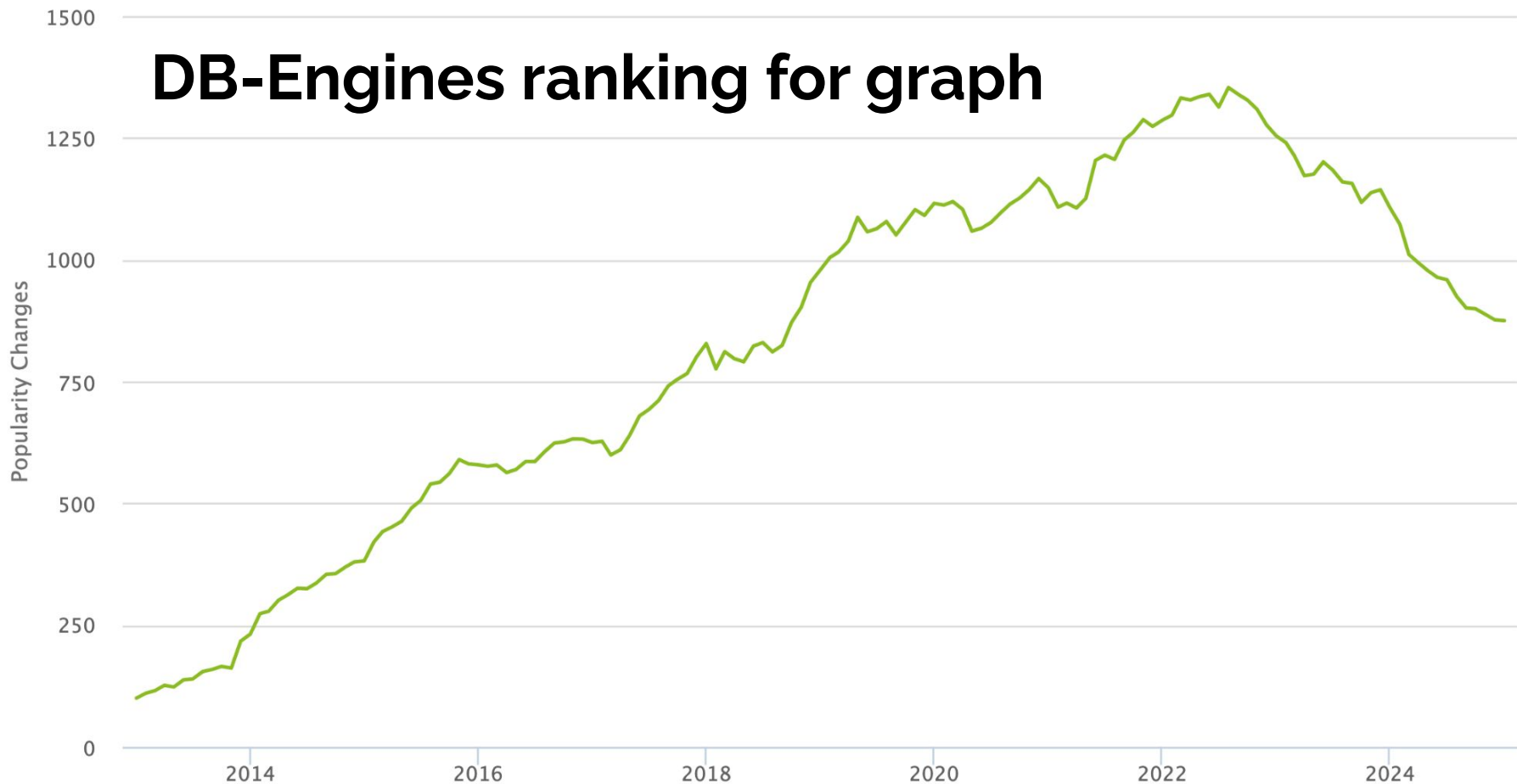
- “X is graph database system” → which category?
- “X doesn't need joins” → unnecessary and confusing conceptual shift
- “graph databases will replace RDBMSs” → this is very unlikely

There are niche use cases, which systems over-optimize for, causing fragmentation

DB-Engines ranking for graph

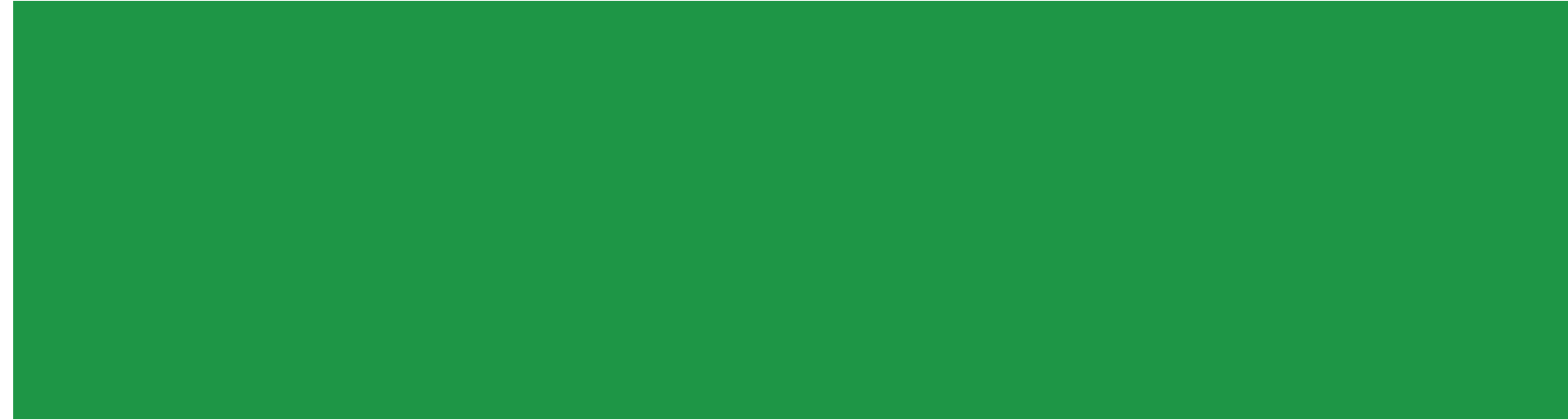


DB-Engines ranking for graph





Summing up



Graph databases <> joins

Graph databases have syntax sugar and optimizations for joins

If you have 10 joins in a query and it doesn't work well, try a graph database:

“MongoDB”

“Postgres”

“Teradata”

Fragmented landscape: graph databases are very specialized

Check licenses and performance results

Sources

Information sheets (all of them contain inaccuracies!):

- Database of databases: <https://dbdb.io/>
- DB-Engines Ranking of Graph DBMS: <https://db-engines.com/en/ranking/graph+dbms>
- Wikipedia page of vendors

Recommended readings / presentations:

- Amine Mhedhbi: [Taming Large Intermediate Results for Joins over Graph-Structured Relations: A System Perspective](#)
- Kùzu blog: <https://blog.kuzudb.com/>
- A eulogy for RedisGraph: <https://www.bloorresearch.com/2023/08/a-eulogy-for-redisgraph/>

Big thanks for discussions to Akon Dey, Amine Mhedhbi and Daniel ten Wolde.

LDBC 

*The graph & RDF
benchmark reference*