UNSW SYDNEY

An Empirical Study on Recent Graph Database Systems

Ran Wang[1], Zhengyi Yang[2], Wenjie Zhang[2], and Xuemin Lin[2]

[1]East China Normal University

[2]University of New South Wales

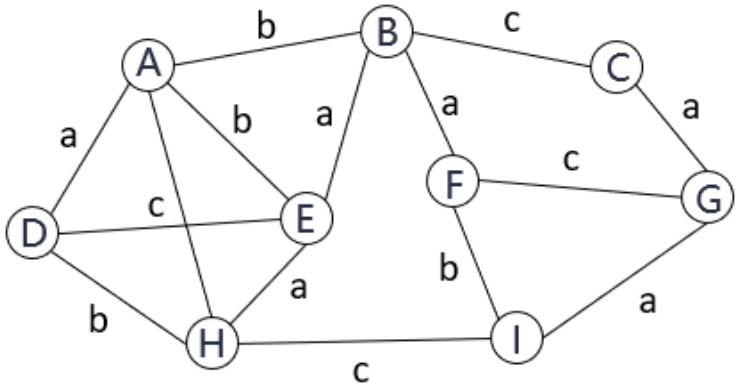
KSEM 2020

Outline

- **Motivation**
- **Overview to Graph Database**
- **Empirical Studies**
- **Conclusion**

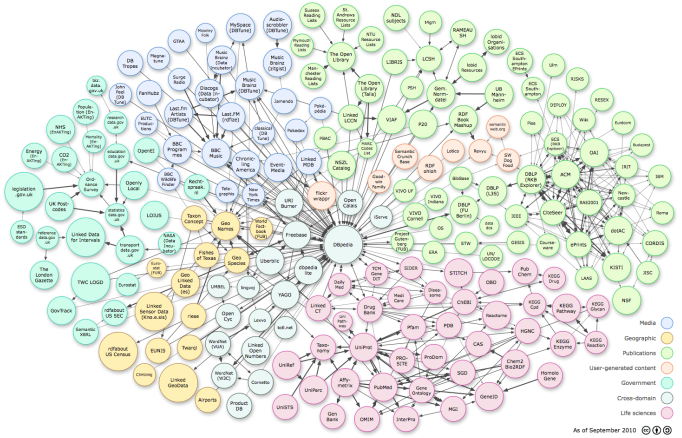
Background

What is Graph?

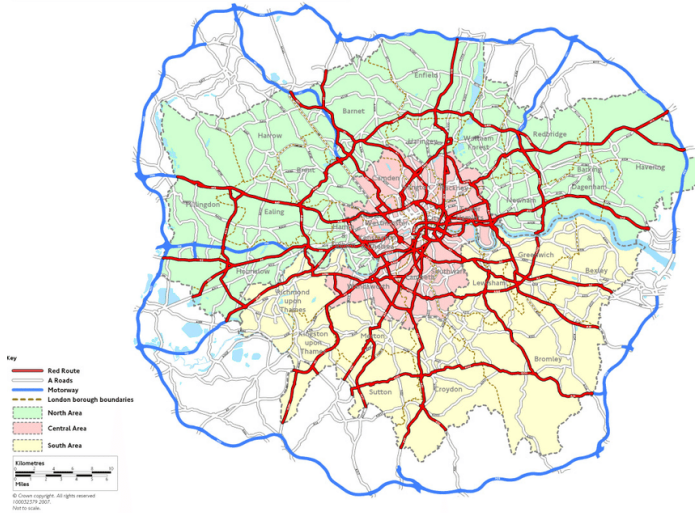


Applications

Social Network



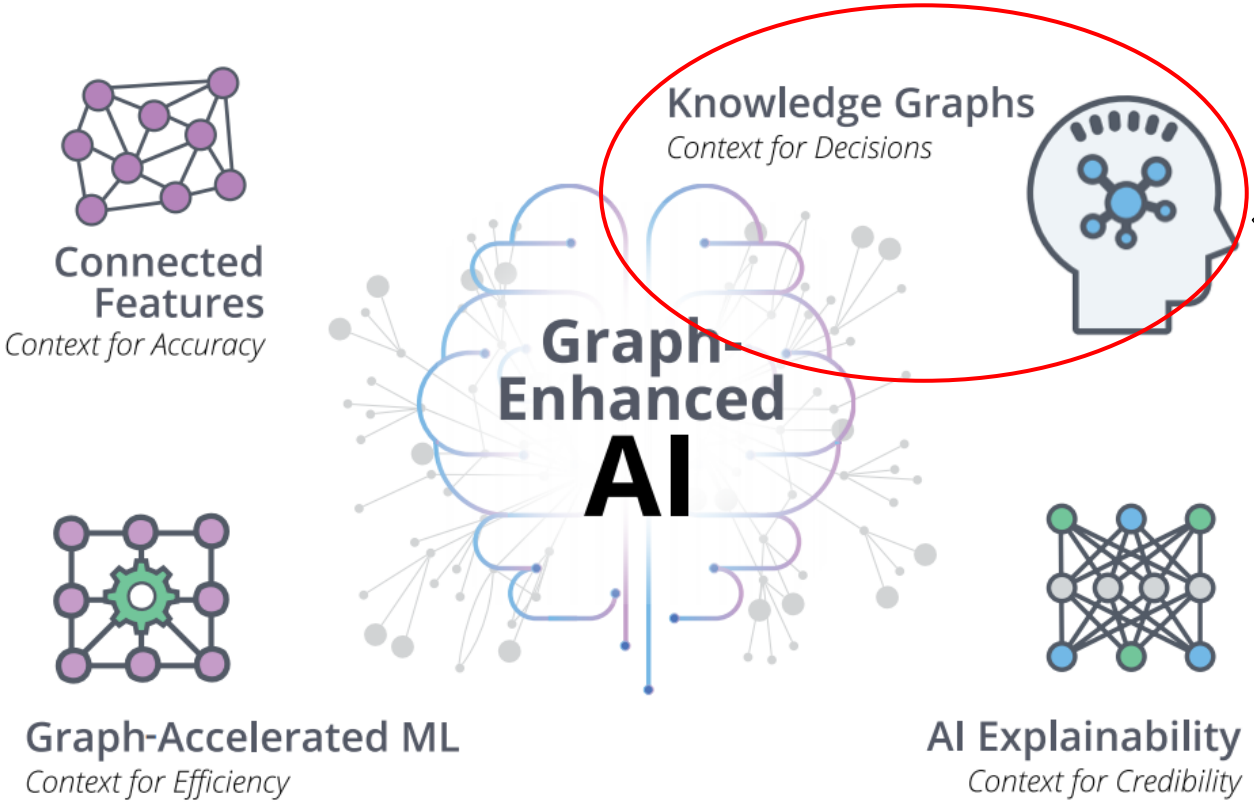
Knowledge Graphs



Road Network

Motivation

Graph Applications in AI & ML

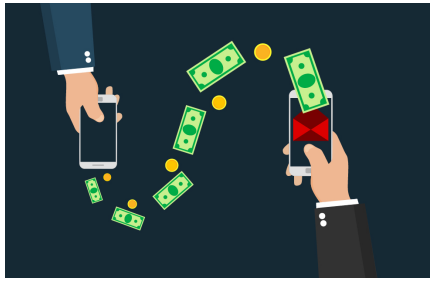


ebay

Recommendation

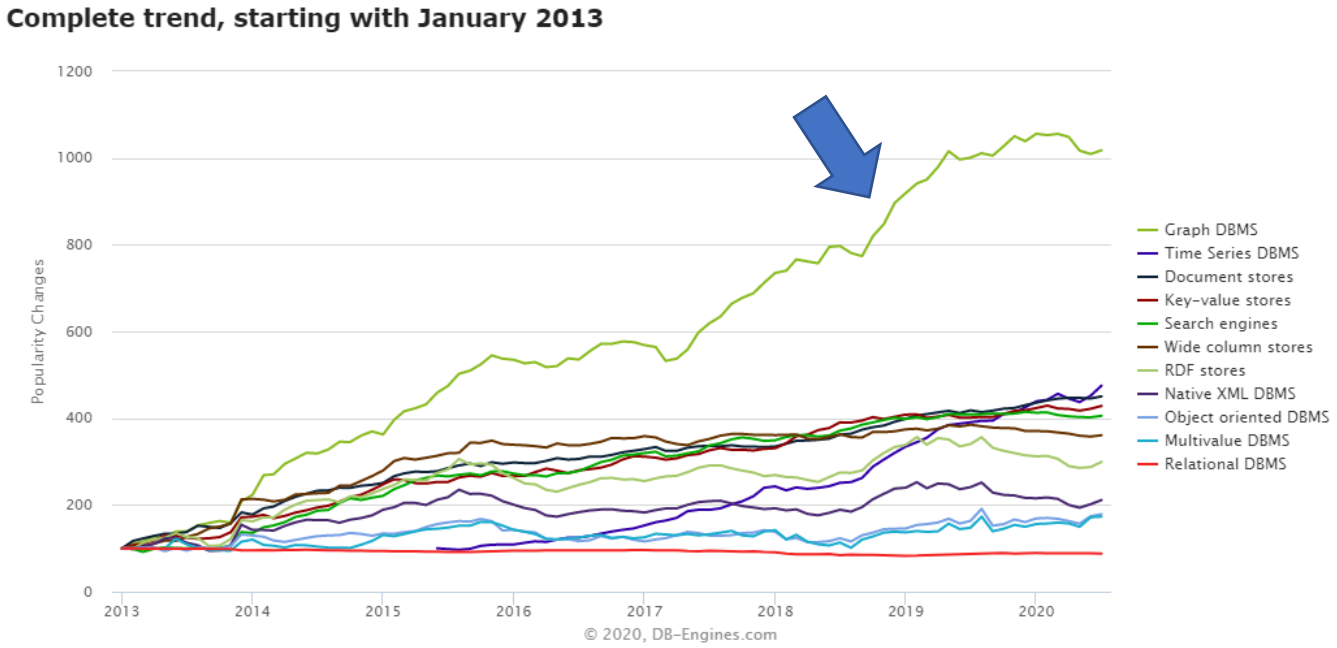
NASA

Lessons Learned database

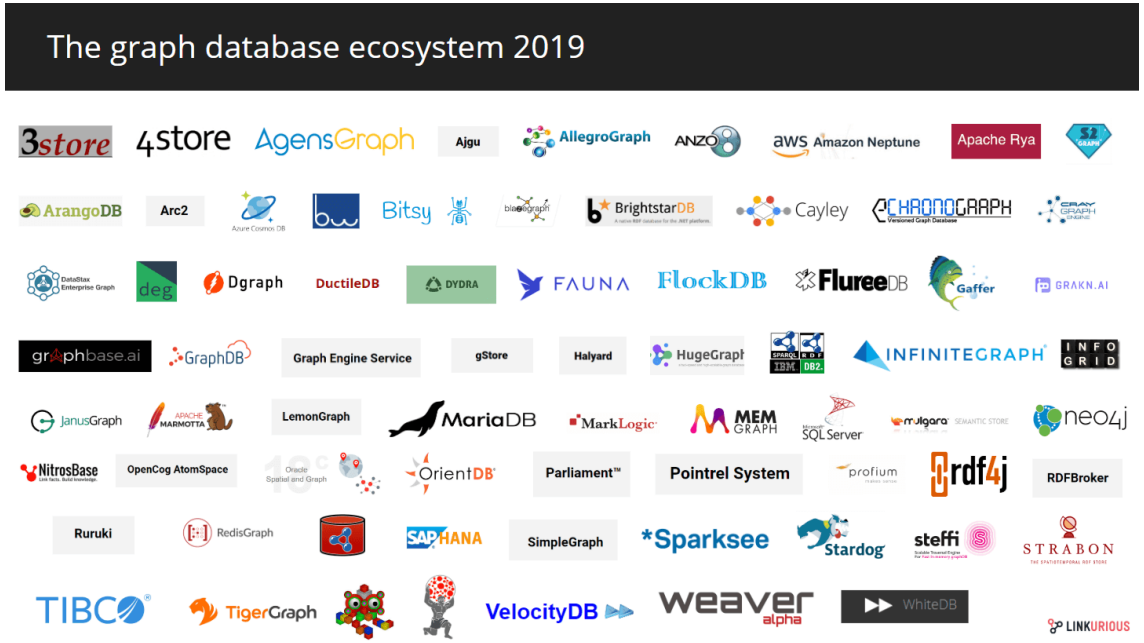


Fraud Detection

Motivation



DBMS popularity trend by database model between 2013 and 2020 – DB-Engine



The graph database landscape in 2019

Motivation

Graph Management Systems

- Graph analytics systems
- Graph databases
  - RDF databases
  - Labeled property graph databases
    - Neo4j
    - JanusGraph
    - ArangoDB
    - TigerGraph
    - AgensGraph
    - LightGraph
    - …

Rank Jul 2020	Rank Jun 2020	Rank Jul 2019	DBMS	Database Model
1.	1.	1	Neo4j	Graph
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model
3.	3.	↑ 4.	ArangoDB	Multi-model
4.	4.	↓ 3.	OrientDB	Multi-model
5.	5.	5.	Virtuoso	Multi-model
6.	6.	↑ 7.	Amazon Neptune	Multi-model
7.	7.	↓ 6.	JanusGraph	Graph
8.	8.	↑ 11.	Dgraph	Graph
9.	↑ 10.	↑ 18.	FaunaDB	Multi-model
10.	↓ 9.	↓ 8.	GraphDB	Multi-model

Motivation

Background

- Popular graph databases have taken a large share of commercial market.
- Many new graph databases are developed in recent years and demonstrate promising performance in their own reports.

WHICH ONE IS THE BEST?

1. *Investigate the features of popular and young graph databases*
2. *Evaluate their performance experimentally*

Motivation

Problems in Existing Benchmark Work

- Lack of experimental exploration

➡ Only list their characteristics.

- Not consider actual and complex business scenarios

➡ Only restrict evaluation on micro operations or small-scale datasets.

- Many young graph databases get little attention

➡ Only limited number of graph databases are studied

Contribution

1. Investigate the market of recent enterprise graph database systems, and present a study of prevalent products.

2. Based on the unified Linked Data Benchmark Council Social Network Benchmark(LDBC SNB), experimentally evaluate popular and young graph databases Neo4j, AgensGraph, TigerGraph and LightGraph.

3. Provide insightful advice on how to select a proper graph database system in different use cases.

Outline

- **Motivation**
- **Overview to Graph Database**
- **Empirical Studies**
- **Conclusion**

Overview

System	Type	Storage Structure	Open Source	Distributed	Transactional	Schema-free	Implementation	Language
Neo4j	Native	Linked List	Yes	No	Yes	Yes	Java	Cypher
JanusGraph	Hybrid	Cassandra/HBase	Yes	Yes	Yes	No	Java	Gremlin
ArangoDB	Hybrid	MMFiles/RocksDB	Yes	Yes	Yes	Yes	C++	AQL
AgensGraph	Hybrid	PostgreSQL	Yes	No	Yes	Yes	C	Cypher,SQL
TigerGraph	Native	Native Engine	No	Yes	Yes	No	C++	GSQL
LightGraph	Native	Native Engine	No	No	Yes	No	C++	Cypher
Nebula	Native	RocksDB	Yes	Yes	No	No	C++	nGQL

- Enterprise graph databases are almost all transactional
- Native databases prefer self-designed storage structures and query languages
- Hybrid databases emerge because of their flexibility
- More and more products target at high scalability

Further Research

Selection Criteria

1. Labeled property graph model ➡ **increasingly popular in industry**
2. Declarative graph query languages ➡ **more user-friendly**
3. Online transaction processing(OLTP) ➡ **wide application**
4. could fully implement query workloads in LDBC_SNB ➡ **full functionality**
5. full license available

Neo4j

✓ The most popular graph database system. Mature community is one of its biggest advantages.

✓ Provide user-friendly interfaces and APIs, and supports many third-party frameworks.

✓ Develop the well-known graph query language Cypher.

✗ *Not support data sharding and cannot scale to very large graphs.*

AgensGraph

✓ New generation multi-model graph database, supporting multiple data models at the same time.

✓ Adopt SQL and Cypher as query languages and can integrate them in one single query.

✗ *Adopt the PostgreSQL RDMS as storage engine, which sometimes restrict the efficiency of loading and querying large graph data.*

✗ *A developing product and cannot support all grammars in Cypher.*

AgensGraph

TigerGraph

✓ One of the rising stars in distributed graph database in recent years, showing strong scalability and great performance.

✓ Develop its own declarative query language GSQL.

✗ GSQL is a stored procedure-like language, and requires extensive knowledge about graphs to write efficient queries.

✗ *A non open source commercial product and not freely available.*

LightGraph

✓ A high-performance graph database, greatly improving the throughput under high loads and enabling queries to be processed with high parallelism.

✓ Support storing and querying billion-scale data in single machine.

✗ *It is still under development. Although LightGraph provides Cypher interface, it still cannot support most grammars.*

✗ *A non open source commercial product and not freely available.*

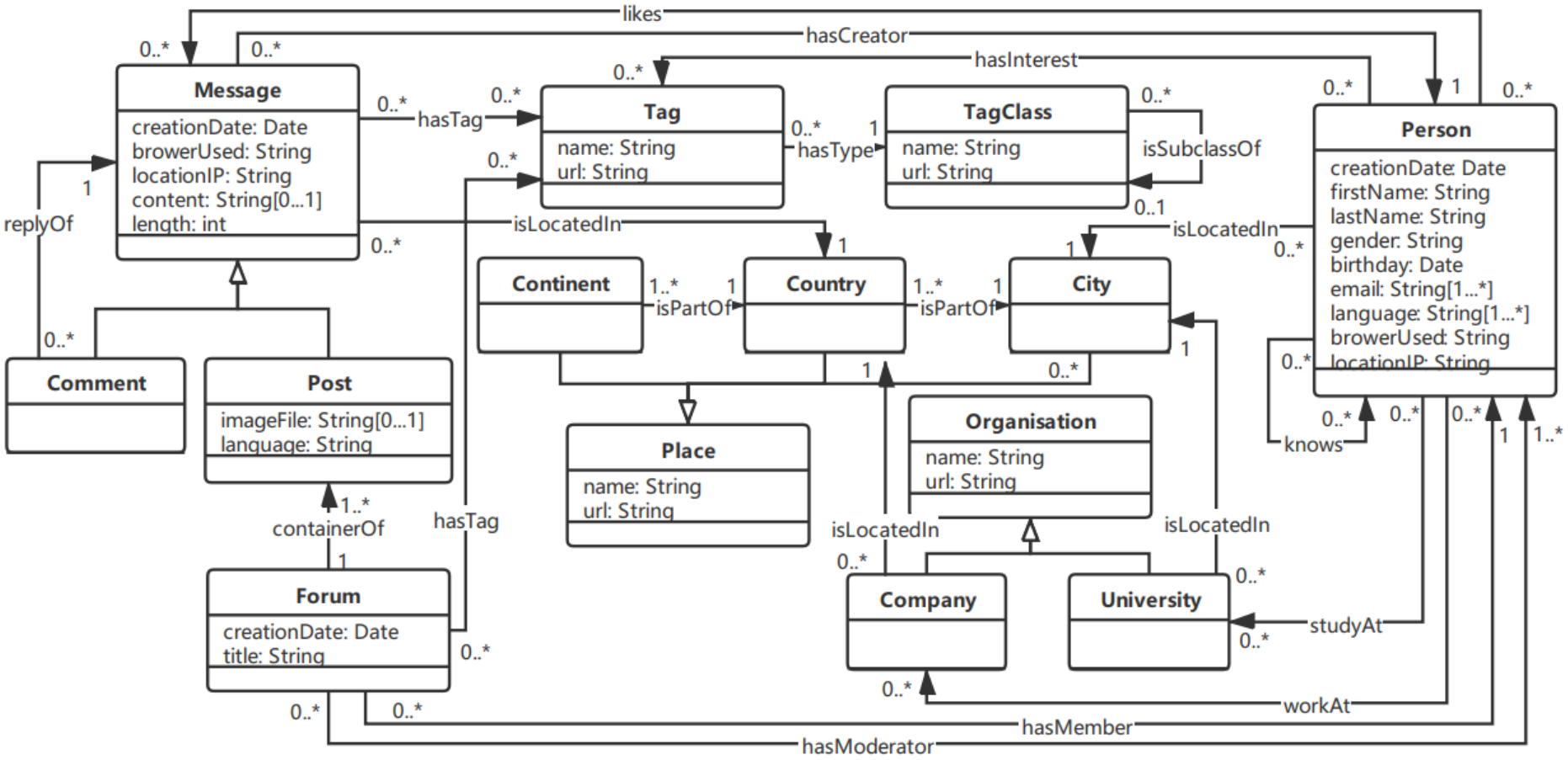
LightGraph

(TuGraph)

Outline

- **Motivation**
- **Overview to Graph Database**
- **Empirical Studies**
- **Conclusion**

Benchmark LDBC_SNB



The schema of LDBC datasets

LDBC The graph & RDF benchmark reference

Query Workloads in LDBC_SNB

✓Interactive Workload
- Transactional update queries
- Simple read-only queries
- Complex read-only queries

✓ Business Intelligence Workload

Workload Type		Abbreviation	Scale	Size
Interactive	Transactional update	IU	Micro	8
	Simple read-only	IS	Micro	7
	Complex read-only	IC	Macro	14
Business intelligence		BI	Macro	25

✗ Graph Algorithms

Setup

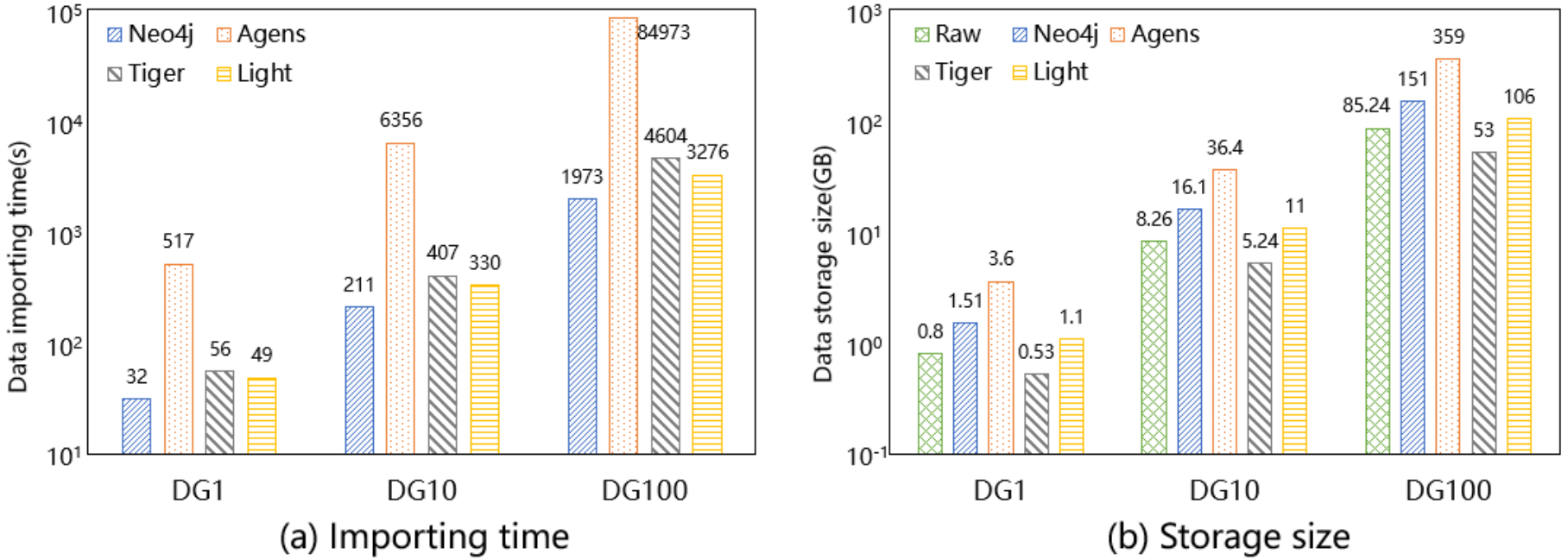
- LDBC_SNB Datasets

Dataset	Scale Factor	|V|(Million)	|E|(Million)	Size(GB)
DG1	1	3.182	17.256	0.798
DG10	10	29.988	176.623	8.257
DG100	100	282.638	1775.514	85.238

- Query Workloads: 54 queries(including micro & macro operations)

- Settings: A machine with two 20-core processors Intel Xeon E5-2680 v2 2.80GHz, 96GB main memory, and 960G NVMe SSD.

Data Importing



- Neo4j performs best in the efficiency of large data importing.
- TigerGraph costs least space to store data.
- LightGraph presents good performance overall.
- AgensGraph shows the worst performance.

Interactive Workload

- Transactional update queries

Table 3. Running time(millisecond) for IU queries.

IU	DG1				DG10				DG100			
	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light
2	4.62	**0.46**	8.14	0.56	4.45	**0.52**	9.81	0.63	4.13	**0.54**	10.36	0.98
4	30.50	8.47	8.55	**1.02**	40.56	7.94	8.52	**1.07**	38.41	9.75	8.90	**1.07**
6	5.02	2.08	8.38	**1.39**	16.10	3.53	8.99	**1.58**	12.01	2.99	9.72	**1.52**
8	1.14	**0.42**	8.18	0.69	1.13	**0.50**	8.45	0.56	1.28	**0.50**	9.40	0.61

- Simple read-only queries

Table 4. Running time(millisecond) for IS queries.

IS	DG1				DG10				DG100			
	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light
1	1.05	1.83	3.47	**0.60**	1.45	1.70	3.27	**0.61**	1.60	1.85	3.65	**0.66**
2	18.49	10.01	10.90	**8.90**	33.11	21.40	**11.09**	15.96	25.80	26.20	**9.91**	16.00
4	1.33	**0.33**	4.01	0.53	1.11	**0.34**	3.74	2.34	0.57	**0.34**	4.06	0.61
6	1.33	13.11	4.66	**0.67**	2.37	14.61	4.7	**0.65**	1.34	15.40	4.41	**0.70**

- Overall: All perform good
- LightGraph: perform best
- AgensGraph: only good at inserting edges
- TigerGraph: not show surprising performance
- Neo4j: perform better than TigerGraph in most cases

Interactive Workload

- Complex read-only queries

Table 5. Running time(second) for IC queries.

IC	DG1				DG10				DG100			
	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light
1	0.60	0.32	**0.03**	0.06	2.36	1.17	**0.12**	0.36	10.22	8.26	**0.56**	2.48
3	2.01	0.35	0.06	**0.01**	24.06	7.95	0.37	**0.05**	616.54	63.82	1.32	**0.37**
6	2.58	0.17	0.09	**0.01**	113.92	0.36	0.31	**0.03**	TO	5.66	0.97	**0.11**
10	0.50	0.71	**0.03**	0.04	2.32	2.93	**0.06**	0.12	9.33	12.41	**0.15**	0.37
12	0.19	2.58	**0.02**	0.04	0.66	4.96	**0.06**	0.15	0.60	55.97	**0.13**	0.16
13	0.01	**0.01**	0.01	0.01	0.03	0.02	**0.01**	0.01	**0.00**	0.03	0.02	0.01
14	340.55	43.34	0.20	**0.01**	424.05	488.61	0.27	**0.02**	63.83	TO	0.31	**0.03**

- Overall: show big differences in performance between graph databases
- TigerGraph & LightGraph: present efficient performance with little difference
- AgensGraph & Neo4j: perform much bad, while good at finding the shortest path, this is because they support the keyword *shortestPath*.

Business Intelligence Workload

BI	DG1				DG10				DG100			
	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light	Neo4j	Agens	Tiger	Light
2	3.36	6.67	0.59	**0.44**	29.16	102.2	**5.27**	9.19	237.6	1388.2	**42.67**	221.8
4	1.58	0.21	**0.02**	0.03	14.50	0.62	0.12	**0.09**	173.3	4.76	**1.19**	3.57
7	375.8	1647.8	0.88	**0.04**	TO	TO	0.88	**0.74**	TO	TO	OOM	**38.63**
8	0.41	1.65	**0.03**	0.08	3.89	6.46	**0.16**	1.00	43.31	69.90	**1.32**	60.07
10	941.2	48.15	0.05	**0.03**	TO	692.2	**0.31**	0.40	TO	TO	**4.11**	33.91
13	0.77	1.39	0.18	**0.12**	5.56	14.36	1.63	**1.07**	46.35	415.5	**13.21**	82.66
16	3.36	TO	**0.49**	0.58	39.97	TO	**4.59**	6.62	429.9	TO	**50.64**	426.1
17	0.41	0.27	0.01	**0.01**	34.65	3.55	**0.03**	0.06	TO	999.1	**0.20**	0.88
18	6.44	352.6	**0.37**	2.95	76.00	TO	**4.71**	58.02	700.1	TO	**52.57**	1742.2
20	4.73	36.53	1.21	**0.84**	44.53	255.1	**14.69**	30.94	732.7	TO	OOM	**290.8**
23	0.18	0.24	**0.02**	0.05	1.55	1.20	**0.06**	0.37	13.44	11.22	**0.51**	50.10
24	16.25	29.96	0.78	**0.66**	191.5	373.2	**7.57**	14.73	TO	TO	**75.71**	1198.7

- Overall: Any system cannot successfully execute all BI queries across all datasets
- TigerGraph & LightGraph: efficient in all cases, while TigerGraph performs better than LightGraph under large-scale datasets
- Neo4j & AgensGraph: timeout in many cases

Overall Evaluation

Summary: Four databases present different performance, and no one can perform best in all scenarios.

Dataset	Good At	Bad At	Usage Experience
Neo4j	➢ data importing ➢ micro queries and small datasets	• high complexity queries • store large-scale datasets	☺ free ☺ user-friendly ☺ complete Cypher
AgensGraph	➢ SQL accompanied workload ➢ simple update & query operations	• process complex queries • manage large datasets	☺ free ☺ for SQL-users
TigerGraph	➢ high complexity queries ➢ manage large datasets	• data importing	☹ not free ☹ query expression
LightGraph	➢ more balanced product ➢ all types of queries	• process business intelligence queries under very large datasets	☹ not free ☹ query expression

Analysis

Potential reasons :

1. The implementation *language differences*: C++(TigerGraph & LightGraph) generally shows a greater performance than Java(Neo4j).

2. Neo4j and AgensGraph are schema-free, while TigerGraph and LightGraph both have *fixed schema*, allowing more optimizations to be done.

3. Commercial products tend to use advanced *algorithms and optimizations.*

4. The *underlaying relational database* of AgensGraph encounters significant extra costs.

Outline

- **Motivation**
- **Overview to Graph Database**
- **Empirical Studies**
- **Conclusion**

Conclusion

- Investigate some graph database systems, then present a further research and evaluation on Neo4j, AgensGraph, TigerGraph, LightGraph.

- Based on the benchmark LDBC_SNB, experiments show that:
  - *LightGraph and TigerGraph have significantly better performance in managing large data and processing high complexity queries.*
  - *Neo4j and AgensGraph give friendly use experience and suitable for micro operations.*

- Future work will extend this study to more graph database products and distributed experiments.

Thank You!

Ran Wang, Zhengyi Yang, Wenjie Zhang, Xuemin Lin

GitHub: https://github.com/UNSW-database/GraphDB-Benchmark