Big Graph Processing Systems

Organisation and **Motivation**

Christopher Spinrath

CNRS – LIRIS – Lyon 1 Université

DISS Master 2025

This presentation is an adaption of slides from Angela Bonifati

Outline

- Motivation
- Part I: Graph Query Paradigms and their Semantics
- Part II: Property Graphs
  - Chapter 1: An introduction to a concrete query language
  - Chapter 2: PG Schema Languages and PG Constraints
  - Chapter 3: Schema Discovery and Property Graph Transformations
- Part III: RDF and Sparql

Organisation

Tentative Schedule

January	06/01	CM + CM	Part I
	13/01	CM + TP	Part II
	20/01	CM + TP	Part II
	27/01	CM + TP	Part II
February	03/02	TP + TP	Part II
	10/02	TP + TP	Part II
	17/02	TP + TP	Part II
	24/02	CM + CM	Part III
	25/02	CM + TP	Part III
March	03/03	CM (Exam)	

- 15 CM + 15 TP
- Please ask (and answer) questions

Organisation

Tentative Schedule

January	06/01	CM + CM	Part I
	13/01	CM + TP	Part II
	20/01	CM + TP	Part II
	27/01	CM + TP	Part II
February	03/02	TP + TP	Part II
	10/02	TP + TP	Part II
	17/02	TP + TP	Part II
	24/02	CM + CM	Part III
	25/02	CM + TP	Part III
March	03/03	CM (Exam)	

- 15 CM + 15 TP
- Please ask (and answer) questions
- Grading
  - 50% Practical lab (TP)
  - 50% Exam (CM)

Hands-on Part – Preparations

- The warm-up project consists of analysing a dataset containing genomic information
- Please install Neo4j and import the dataset before the first hands-on lesson

Hands-on Part – Preparations

- The warm-up project consists of analysing a dataset containing genomic information
- Please install Neo4j and import the dataset before the first hands-on lesson

1. Install Neo4j
   - Make sure you are using version 5.26 or newer
   - Make sure you have enough memory, at least 20GB
   - A simple option is to install Neo4j Desktop by follwowing the official instructions: `https://neo4j.com/docs/desktop-manual/current/installation/`
2. Download and import the database
   - Download the database dump file from `https://partage.liris.cnrs.fr/index.php/s/LoEtp24fk38P6n5`
   - Import the dump file by following the official instructions: `https://neo4j.com/docs/desktop-manual/current/operations/create-from-dump/`

Everything is Data

[https://www.acronis.com/en-us/blog/posts/data-everything-8-noble-truths/]

A New Era of Science

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data.

Welcome to the ~~Peta~~Zettabyte Age.

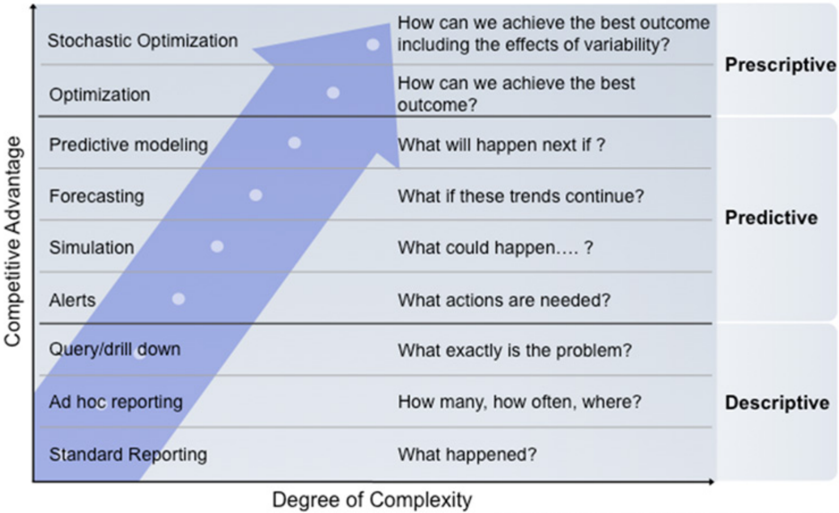
New Realities

- Everything is data
- Rise of data-driven culture
- High-performant data analytics
- Exploit sophisticated statistical methods

How do we structure/implement/live with this trend?

The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
WIRED 16.07
TMI | JUL.2008
How do we find DARK MATTER?
Do we need THEORY?
Is Earth safe from ASTEROIDS?
Where will TERRORISTS strike next?
How can you win any LAWSUIT?
Who will be the next PRESIDENT?
How will we grow enough FOOD?
How can we protect BUSINESS from risk?
The End of Science
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. Welcome to the Petabyte Age.

Level of Analytics



Stochastic Optimization	How can we achieve the best outcome including the effects of variability?	Prescriptive
Optimization	How can we achieve the best outcome?	
Predictive modeling	What will happen next if ?	Predictive
Forecasting	What if these trends continue?	
Simulation	What could happen.... ?	
Alerts	What actions are needed?	
Query/drill down	What exactly is the problem?	Descriptive
Ad hoc reporting	How many, how often, where?	
Standard Reporting	What happened?	

Based on: Competing on Analytics, Davenport and Harris, 2007

Focus of Interest

Properties of Entities

- Captured/measured values
- What are the sales figures/temperatures/etc.?
- Multidimensional data/time series/matrices

Connections Between Entities

- Network structure
- What do the friends of your customers buy?
- Graph data

Connectedness

Connections Manifest in Many Different Ways, Facets, Values, Scopes, and Scales

Cause: interaction, proximity, information, relationship, affiliation, ownership, allegiance, force, repulsion, …

Nature: social, cultural, economical, physical, chemical, …

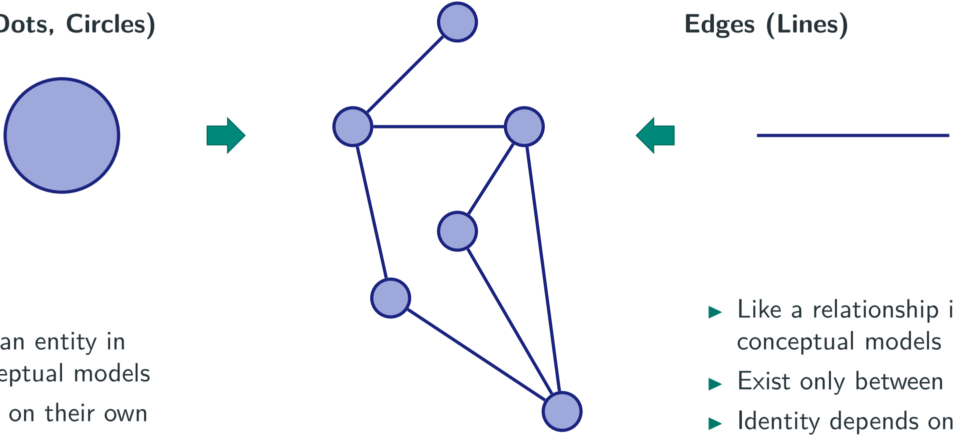
Quality: existential, essential, inconsequential, or indiscernible; persistent or temporary; positive, neutral, or negative

Scale: from interacting cells, such as neurons in our brain, to high-scale entities, such as families, groups, clubs, companies, organizations, states, nations, … or stars and galaxies

```
(graphs)-[:ARE]->(everywhere)
```

Graph Building Blocks

Nodes (Dots, Circles)



Edges (Lines)

- Like an entity in conceptual models
- Exist on their own
- Have an object identity

- Like a relationship in conceptual models
- Exist only between nodes
- Identity depends on the nodes they connect

Graphs as Unifying Abstractions

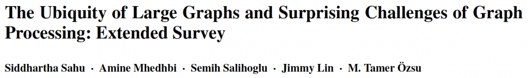
- Graphs are natural abstractions
  for representing interconnected objects when encoding, explaining and predicting real-world and digital-world phenomena.
- Graphs are underpinning several data management ecosystems,
  in societal, scientific, RDF, product and digital domains.
- There is no unique killer application for graphs,
  but several exist.
- Nevertheless, the data models, query languages and system requirements
  needed for graphs are constantly evolving.

CONTRIBUTED ARTICLES

The Future Is Big Graphs: A Community View on Graph Processing Systems

By Sherif Sakr, Angela Bonifati, Hannes Voigt, Alexandru Iosup, Khaled Ammar, Renzo Angles, Walid Aref, Marcelo Arenas, Maciej Besta, Peter A. Boncz, Khuzaima Daudjee, Emanuele Della Valle, Stefania Dumbrava, Olaf Hartig, Bernhard Haslhofer, Tim Hegeman, Jan Hidders, Katja Hose, Adriana Iamnitchi, Vasiliki Kalavri, Hugo Kapp, Wim Martens, M. Tamer Özsu, Eric Peukert, Stefan Plantikow, Mohamed Ragab, Matei R. Ripeanu, Semih Salihoglu, Christian Schulz, Petra Selmer, Juan F. Sequeda, Joshua Shinavier
Communications of the ACM, September 2021, Vol. 64 No. 9, Pages 62-71
10.1145/3434642
Comments

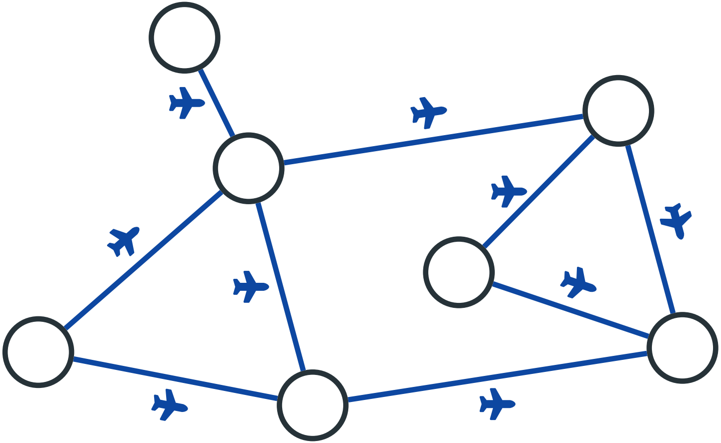
ARTICLE CONTENTS:
Introduction
Key Insights
Abstractions
Ecosystems
Performance



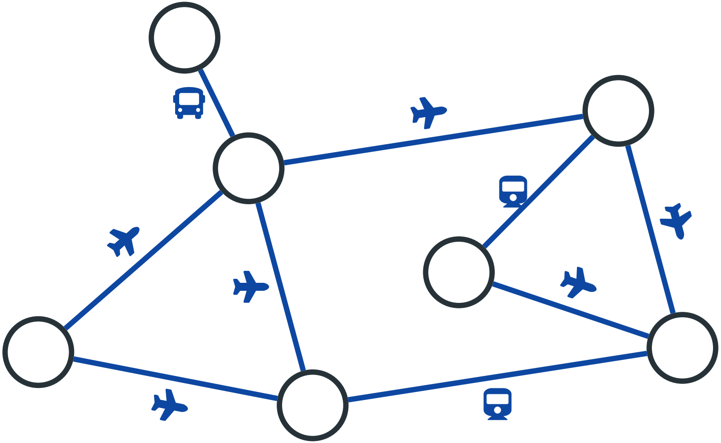
The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing: Extended Survey

Siddhartha Sahu · Amine Mhedhbi · Semih Salihoglu · Jimmy Lin · M. Tamer Özsu

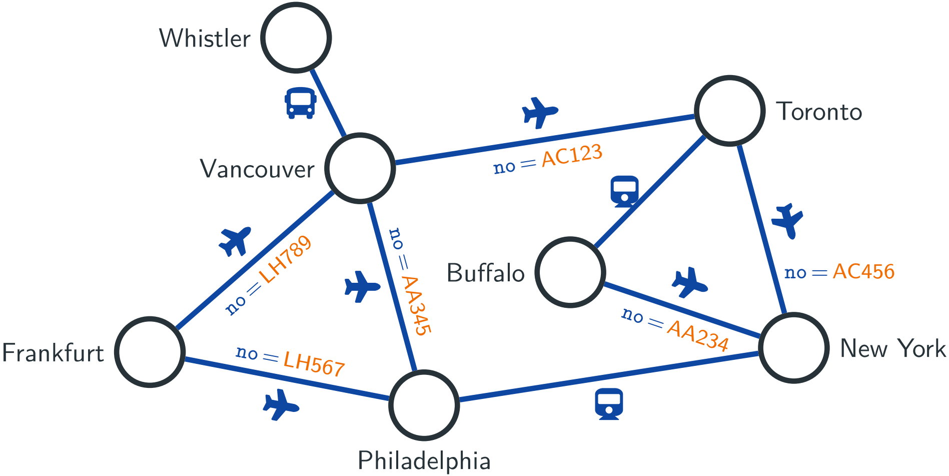
Graph Data – Transportation Network



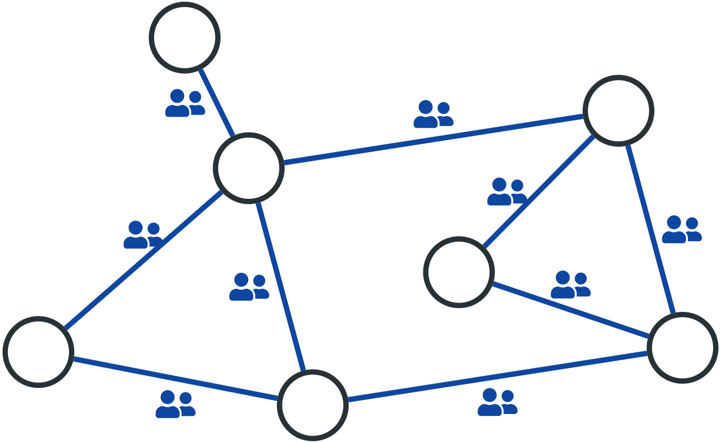
Graph Data – Transportation Network



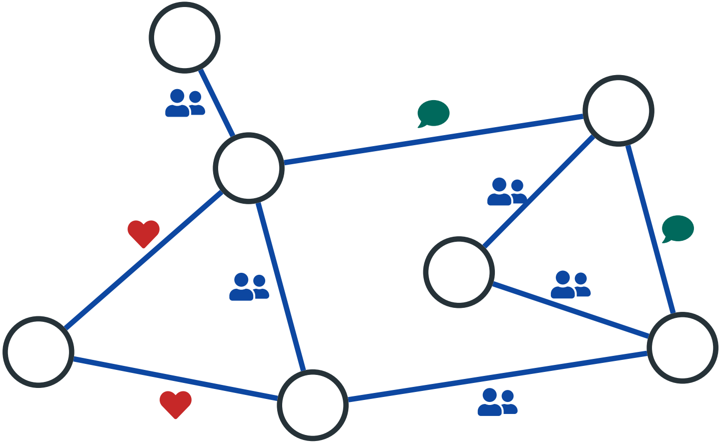
Graph Data – Transportation Network



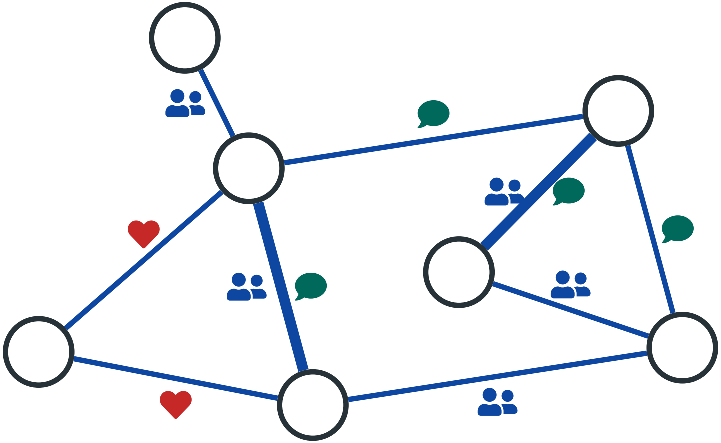
Graph Data – Social Network



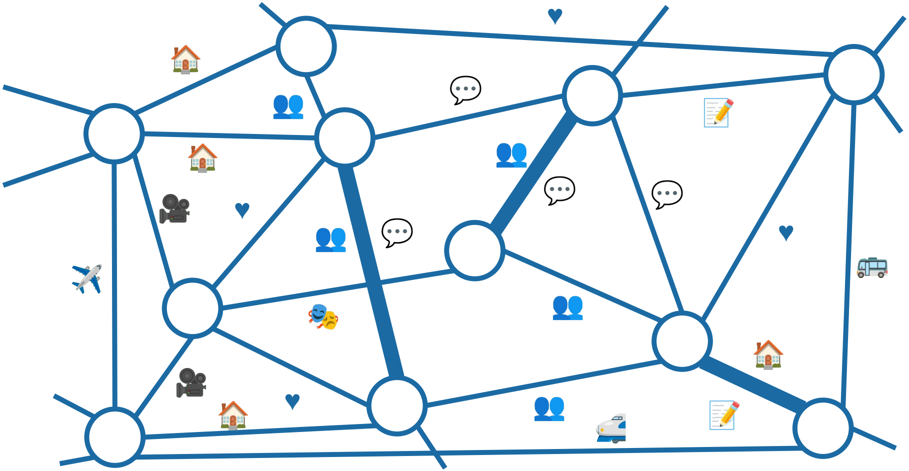
Graph Data – Social Network



Graph Data – Social Network



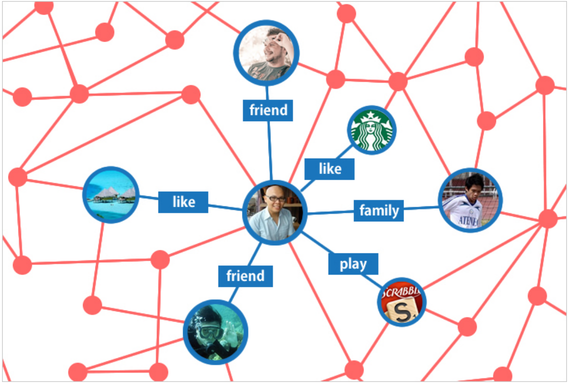
Graph Data – Heterogeneity

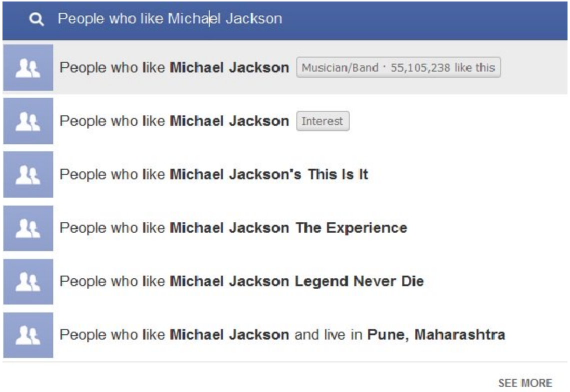


Structured Search

Example (Facebook Graph Search)

- Finding subgraph structures
- Very natural way of formulating queries





People who like Michael Jackson

People who like **Michael Jackson** Musician/Band · 55,105,238 like this

People who like **Michael Jackson** Interest

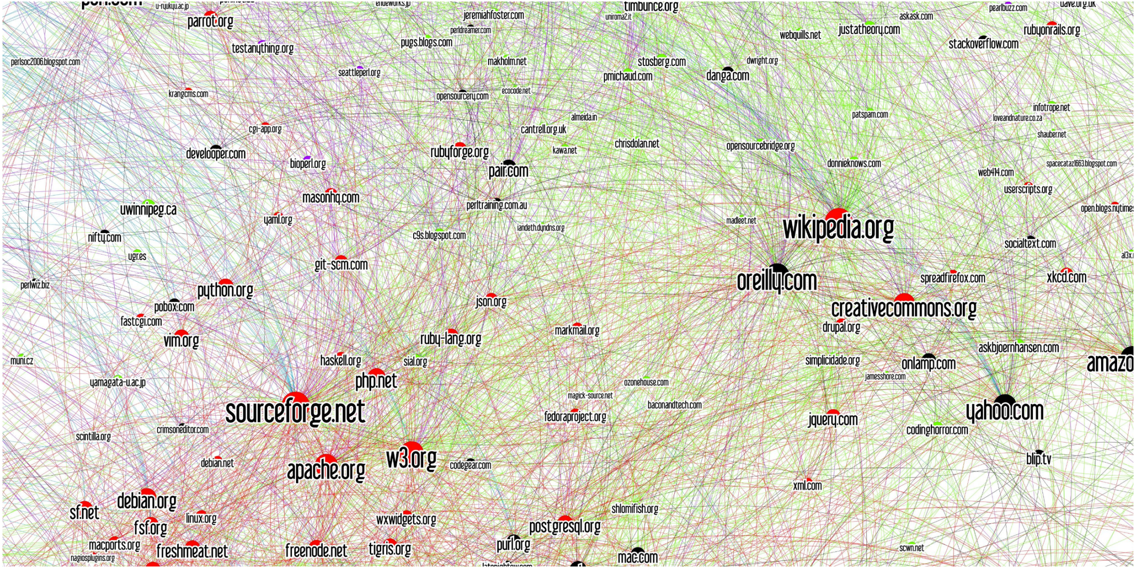
People who like **Michael Jackson's This Is It**

People who like **Michael Jackson The Experience**

People who like **Michael Jackson Legend Never Die**

People who like **Michael Jackson** and live in **Pune, Maharashtra**

Web Graph

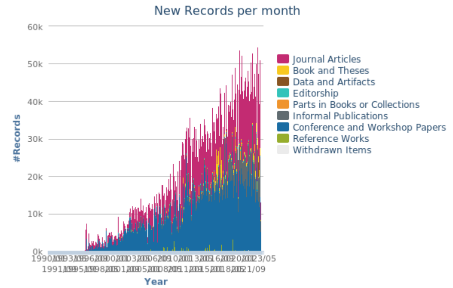


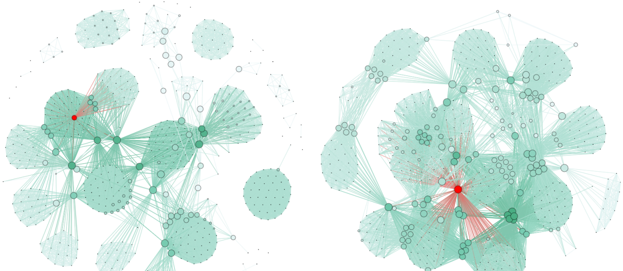
[http://cpan-explorer.org/2009/06/12/web-community-graph/]

Bibliography

Example (DBLP)

- Open bibliographic information on major computer science journals and proceedings
- >6.5 million publication
- >46000 new publication per month
- >1.7 million authors





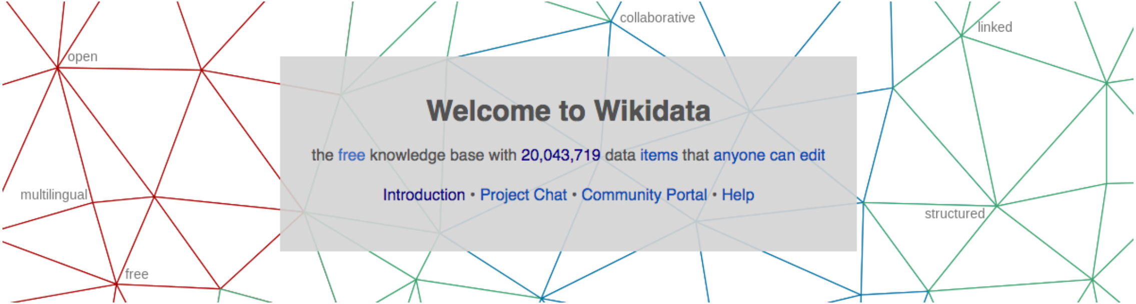
[https://dblp.org/statistics/newrecordspermonth.html]

Knowledge Graphs – Wikidata

Example (Graph to capture world-knowledge)

- Open knowledge base that can be read and edited by humans and machines
- Structured data of Wikipedia, Wikivoyage, Wikisource, etc.

WIKIDATA



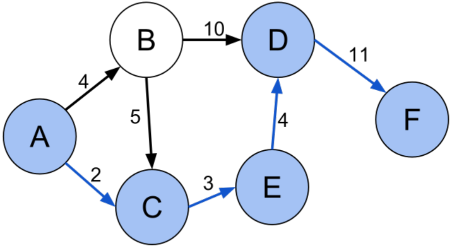
Welcome to Wikidata

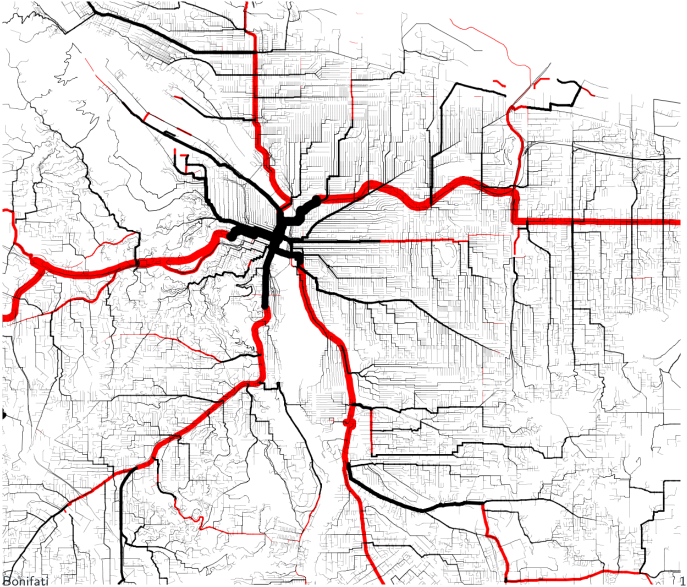
the free knowledge base with 20,043,719 data items that anyone can edit

Shortest Paths

Example (Transportation Network)

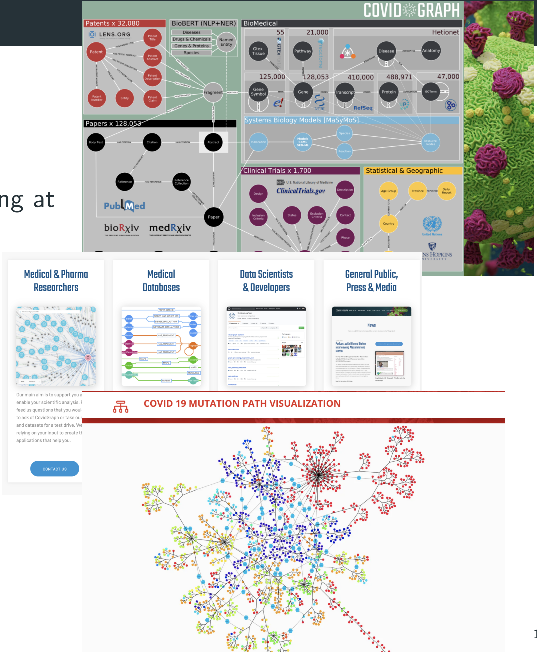
- Only hop-distances (#edges)
  - in friendship network
- With weights (distance, travel time, etc.)
  - in road networks, transportation connections





A Plethora of Applications

- Among which, the `covidgraph.org` initiative aiming at building the Covid19 knowledge graph
  - Collecting patents, publications about the human coronaviruses
  - Biomedical data (genomics and omics)
  - Experimental data about clinical trials
  - Key demographic indicators
- Practical use case in many data-oriented tasks
  - Property graph schema discovery
  - Threshold queries in Theory and in the Wild


COVID GRAPH
Patents x 32,080
BioBERT (NLP+NER)
BioMedical
Papers x 128,053
Systems Biology Models [MaSyMoS]
Clinical Trials x 1,700
Statistical & Geographic
Medical & Pharma Researchers
Medical Databases
Data Scientists & Developers
General Public, Press & Media
CONTACT US
COVID 19 MUTATION PATH VISUALIZATION

Our main aim is to support you a
enable your scientific analysis. F
feed us questions that you woul
to ask of CovidGraph or take our
and datasets for a test drive. We
relying on your input to create th
applications that help you.

Threshold graph queries on the Covid19 graph

- Find each country that does not have three reports for some age group

```
MATCH (c:Country)
         -[e:CURRENT_FEMALE | CURRENT_MALE | CURRENT_TOTAL]->
      (a:AgeGroup)
WITH c, a, COUNT(type (e)) AS ecount
WHERE ecount < 3 RETURN c, a
```

Adapted from

[Bon22a] Angela Bonifati, Stefania Dumbrava, George Fletcher, Jan Hidders, Matthias Hofer, Wim Martens, Filip Murlak, Joshua Shinavier, Slawek Staworko, Dominik Tomaszuk: Threshold Queries in Theory and in the Wild. Proc. VLDB Endow. 15(5): 1105-1118 (2022)

Threshold graph queries on the Covid19 graph

- Find each country that does not have three reports for some age group

```
MATCH (c:Country)
        -[e:CURRENT_FEMALE | CURRENT_MALE | CURRENT_TOTAL]->
     (a:AgeGroup)
WITH c, a, COUNT(type (e)) AS ecount
WHERE ecount < 3 RETURN c, a
```

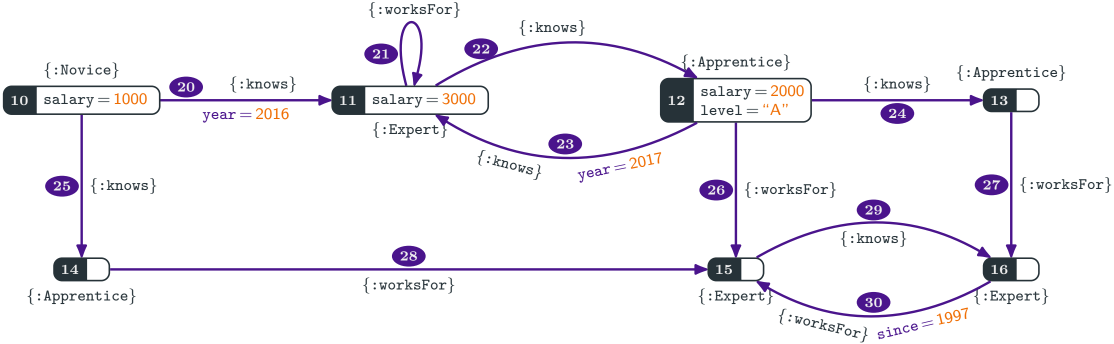
- Find each protein that has more than 43917 associated geneontology terms

```
MATCH (p:Protein )-[:MAPS]->*()-[:HAS_ASSOCIATION]->()
     (()-[:IS_A]->*()|()-[:PART_OF]->*())(t:GOTerm)
WITH p, COUNT (DISTINCT t) AS count_go
WHERE count_go > 43917 RETURN p
```

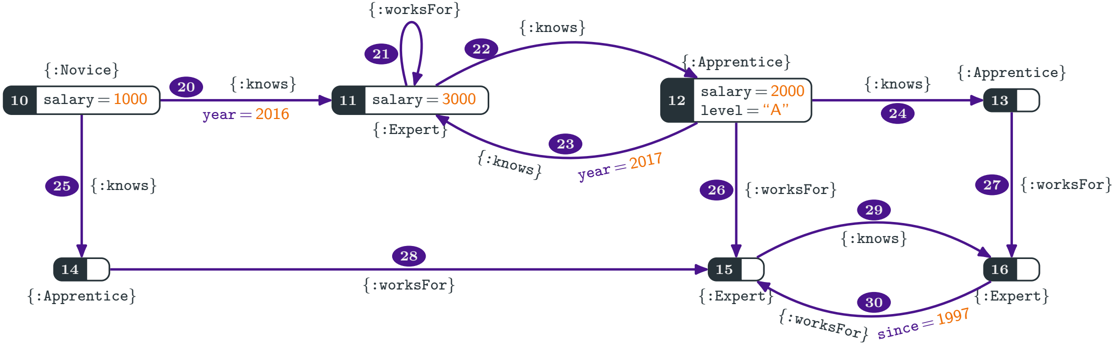
Adapted from

[Bon22a] Angela Bonifati, Stefania Dumbrava, George Fletcher, Jan Hidders, Matthias Hofer, Wim Martens, Filip Murlak, Joshua Shinavier, Slawek Staworko, Dominik Tomaszuk: Threshold Queries in Theory and in the Wild. Proc. VLDB Endow. 15(5): 1105-1118 (2022)

Property Graphs – Example



Property Graphs – Example



Nodes and edges have

- IDs: 10, 11, ...
- labels, e.g., :Novice, :Apprentice, :worksFor, ...
- properties, e.g. "salary = 3000", "year = 2016", ...

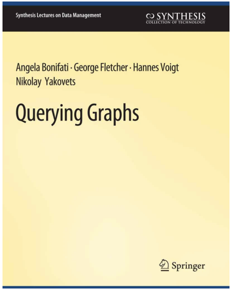
Property Graphs: A formal definition

Assume pairwise disjoint sets of
$\mathcal{O}$ (objects), $\mathcal{L}$ (labels), $\mathcal{K}$ (property keys), and $\mathcal{N}$ (values)

Definition

A property graph is a structure $(V, E, \eta, \lambda, \vartheta)$ where

- $V \subseteq \mathcal{O}$ is a finite set of vertices,
- $E \subseteq \mathcal{O}$ is a finite set of edges,
- $\eta\colon E \to V \times V$ assigns an ordered pair of vertices to each edge,
- $\lambda\colon V \cup E \to \mathcal{P}(\mathcal{L})$ assigns a finite set of labels to each vertex and edge,
- $\vartheta\colon (V \cup E) \times \mathcal{K} \rightharpoonup \mathcal{N}$ assigns values for properties to vertices and edges.



Synthesis Lectures on Data Management

SYNTHESIS COLLECTION OF TECHNOLOGY

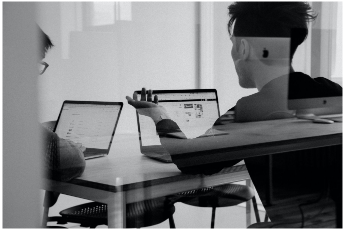
Angela Bonifati · George Fletcher · Hannes Voigt
Nikolay Yakovets

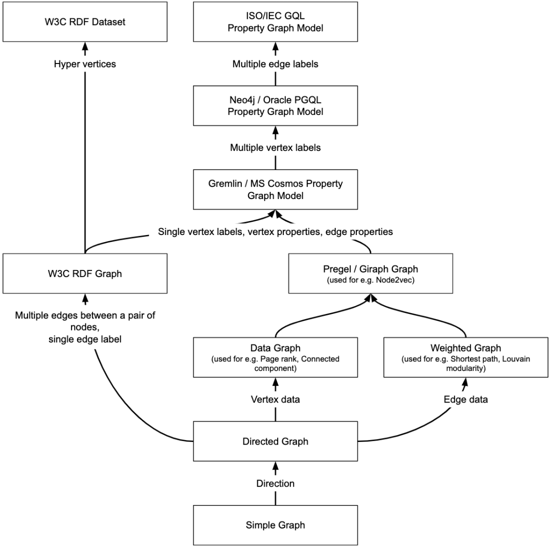
Querying Graphs

Springer

A Lattice of Graph Data Models

- A data model per use case
- How expressive and human-friendly is a data model?
- Need of making different data models interoperable via mappings or direct translations





W3C RDF Dataset
ISO/IEC GQL Property Graph Model
Hyper vertices
Multiple edge labels
Neo4j / Oracle PGQL Property Graph Model
Multiple vertex labels
Gremlin / MS Cosmos Property Graph Model
Single vertex labels, vertex properties, edge properties
W3C RDF Graph
Pregel / Giraph Graph (used for e.g. Node2vec)
Multiple edges between a pair of nodes, single edge label
Data Graph (used for e.g. Page rank, Connected component)
Weighted Graph (used for e.g. Shortest path, Louvain modularity)
Vertex data
Edge data
Directed Graph
Direction
Simple Graph