Knowledge Graphs

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, Antoine Zimmermann

KnowGraphs Colloquium

November 27, 2020

Introduction

- Tutorial paper
  - Comprehensive introduction to multiple research areas
- Focus of this presentation
  - Data Graphs

Data Graphs

Data Graphs

Overview

- Graph Models
  - Directed edge-labelled graph
  - Graph dataset
  - Property graph
- Querying
  - Query languages (e.g., SPARQL, Cypher, etc.)
  - Graph patterns
  - Complex graph patterns
  - Navigational graph patterns

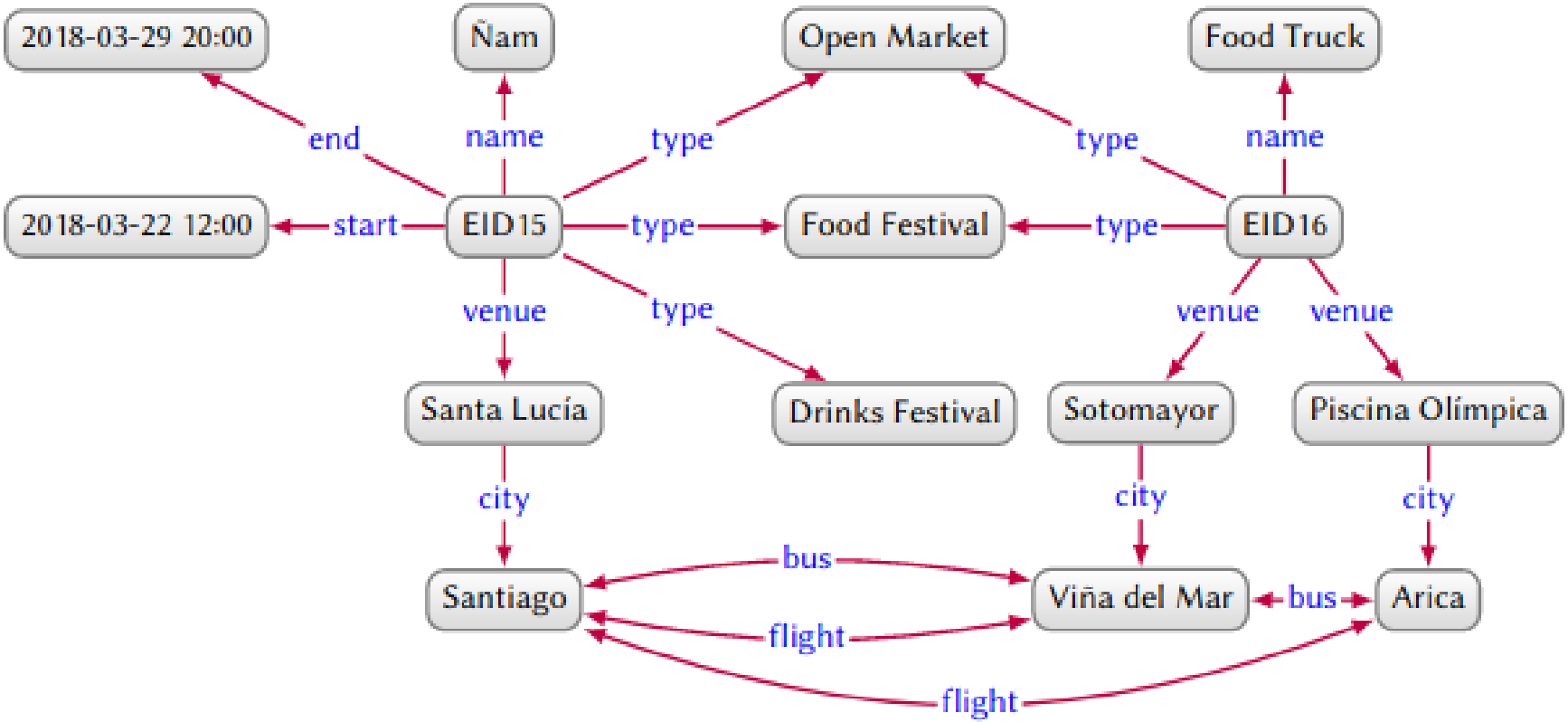
Definition

Let Con be a an infinite set of constants. A directed edge-labelled graph is a tuple $G := (V, E, L)$, where $V \subseteq Con$ is a set of nodes, $L \subseteq Con$ is a set of labels and $E \subseteq V \times L \times V$ is a set of edges.

- V: set of nodes representing entities
  - e.g., *Arica*, *Santiago*
- E: set of edges representing relations between entities
  - e.g., (*Arica*, *flight*, *Santiago*)

Data Graphs

Graph Models - Directed Edge-Labelled Graph (cont'd)



Data Graphs

Graph Models - Directed Edge-Labelled Graph (cont'd)

- **R**esource **D**escription **F**ramework
  - W3C recommendation
  - **I**nternationalized **R**esource **I**dentifiers
    - Global identification of entities on the Web
  - Literals
    - Representation of datatypes (e.g., integers, dates, etc.)
  - Blank nodes
    - Nodes without identifiers

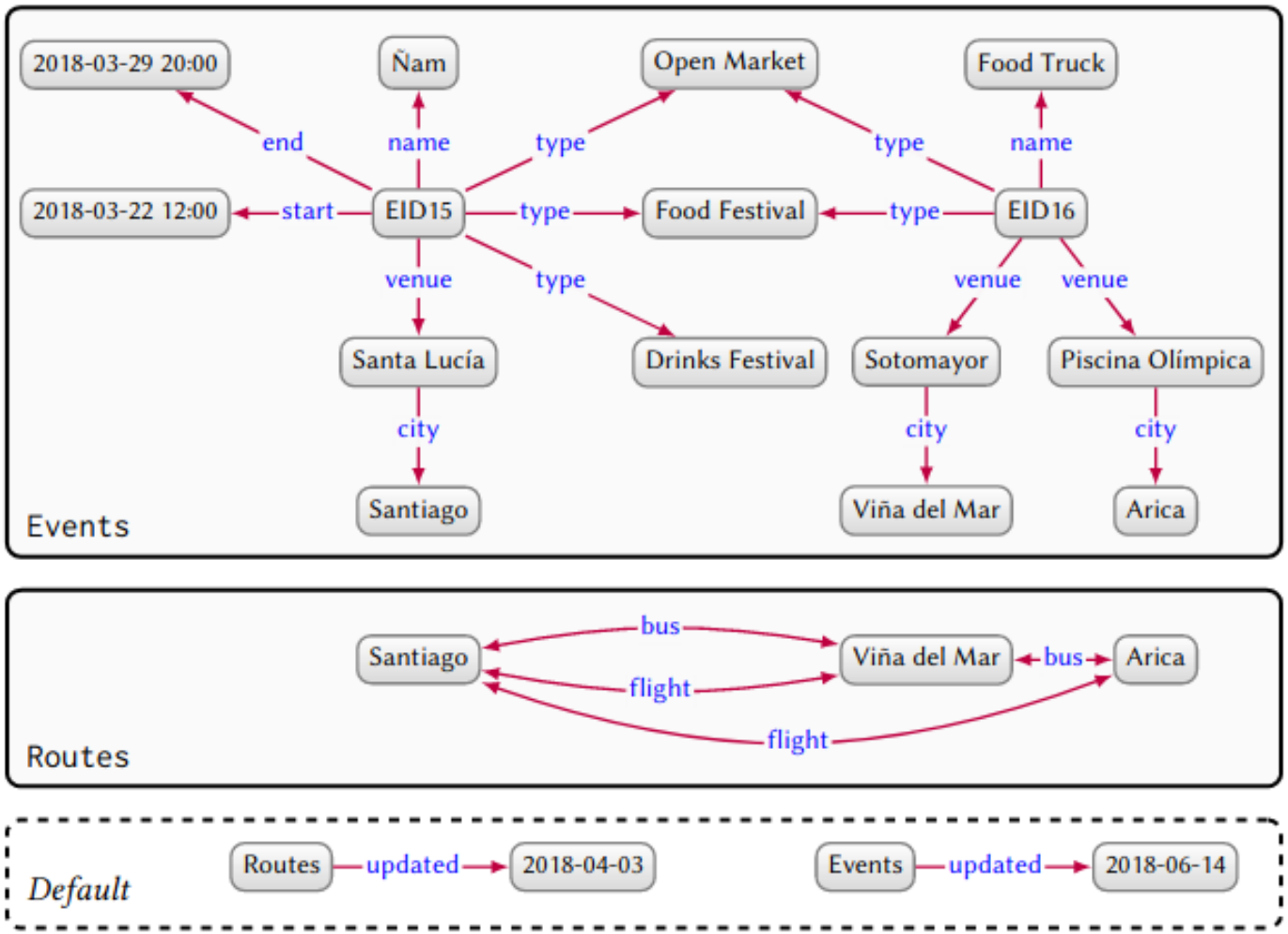
Definition

A named graph is a pair (n, G) where G is a directed edge-labelled graph, and $n \in Con$ is a graph name. A graph dataset is a pair $D := (G_D, N)$ where G_D is a directed edge-labelled graph called the default graph and N is either the empty set or a set of named graphs $\{(n_1, G_1), \ldots, (n_k, G_k)\}$ $(k > 0)$ such that $n_i = n_j$ if and only if $i = j$ $(1 \leq i \leq k, 1 \leq j \leq k)$.

- Multiple graphs from different sources
  - Example: Linked Data on the web
- Default graph
  - Does not have an identifier
- Named graphs
  - Each named graph is associated with an identifier

Data Graphs

Graph Models - Graph Dataset

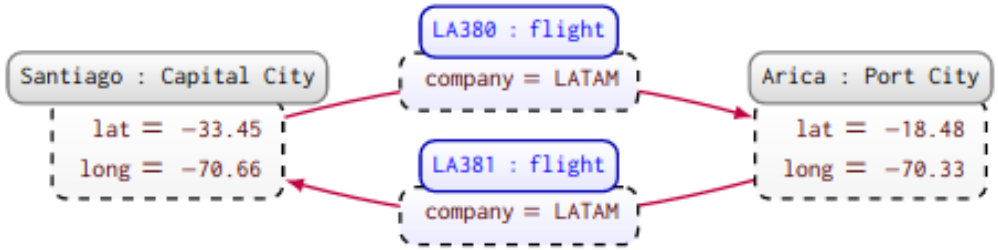


Definition

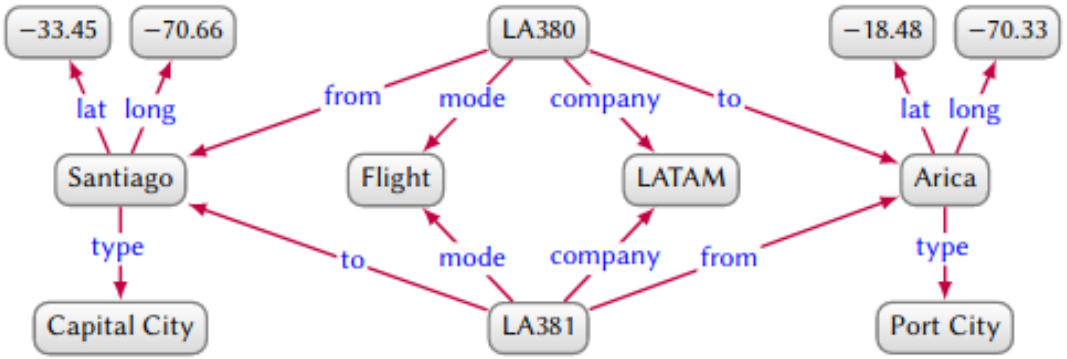
A property graph is a tuple $G := (V, E, L, P, U, e, l, p)$, where $V \subseteq Con$ is a set of node ids, $E \subseteq Con$ is a set of edge ids, $L \subseteq Con$ is a set of labels, $P \subseteq Con$ is a set of properties, $U \subseteq Con$ is a set of values, $e : E \to V \times V$ maps an edge id to a pair of node ids, $l : V \cup E \to 2^L$ maps a node or edge id to a set of labels, and $p : V \cup E \to 2^{P \times U}$ maps a node or edge id to a set of property–value pairs.

- l: assigns labels to nodes and edges
- p: assigns property-value pairs to nodes and edges
- Flexible data model

- Property graph

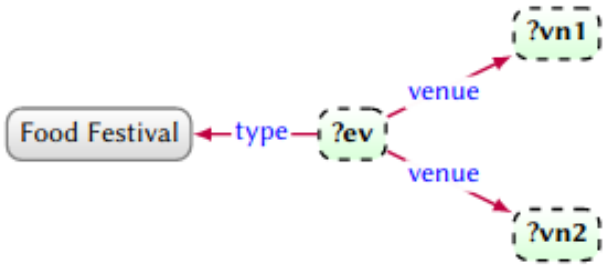


- Directed edge-labelled graph



- RDF
  - SPARQL
- Property graphs
  - Cypher
  - Gremlin
  - G-Core
- Different models - similar underlying characteristics
  - Graph patterns
  - Complex graph patterns
  - Navigational graph patterns

- Core of every graph query language
- Follow the same model as the target data graph
  - Terms of the graph (e.g., nodes) can be variables
- Mappings from variables to constants
  - Table of results
- Semantics
  - Homomorphism-based: **multiple** variables can be mapped to the same term
  - Isomorphism-based: variables are mapped to **unique** terms

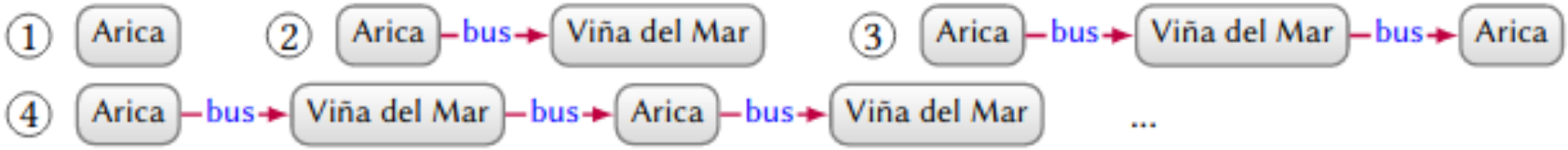


?ev	?vn1	?vn2
EID16	Piscina Olímpica	Sotomayor
EID16	Sotomayor	Piscina Olímpica
EID16	Piscina Olímpica	Piscina Olímpica
EID16	Sotomayor	Sotomayor
EID15	Santa Lucía	Santa Lucía

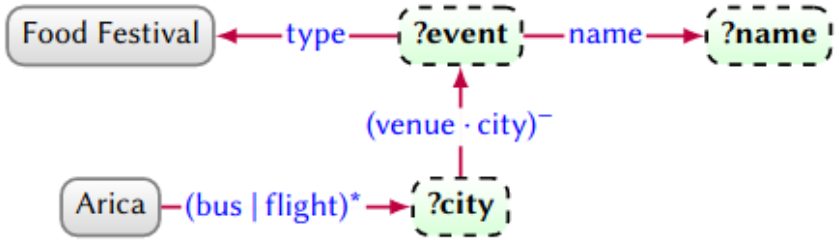
- Combination of result tables using relational algebra
- Operators
  - Unary (e.g., projection (π), etc.)
  - Binary (e.g., union ($\cup$), join ($\bowtie$), etc.)
- Semantics:
  - Bag: allow duplicate results
  - Set: remove duplicate results (DISTINCT keyword)

- Combination of graph patterns and regular path queries
- Regular path query: (x, r, y)
  - r: path expression, arbitrary-length paths between two nodes
    - r, base path expression
    - r^*, zero-or-more paths
    - r^-, inverse path
    - $r_1|r_2$, disjunction of paths
    - $r_1 \bullet r_2$, conjunction of paths

- Example, regular path query: (Arica, bus^*, ?city)



- Example, navigational graph pattern



?event	?name	?city
EID15	Ñam	Santiago
EID16	Food Truck	Arica
EID16	Food Truck	Viña del Mar

Schema, Identity, Context

Schema, Identity, Context

Overview

- From data graphs to knowledge graphs
- Schema
  - Prescribes a high-level structure and/or semantics
  - Semantic
  - Validating
  - Emergent
- Identity
  - Node disambiguation
- Context
  - Scope of truth

Schema, Identity, Context

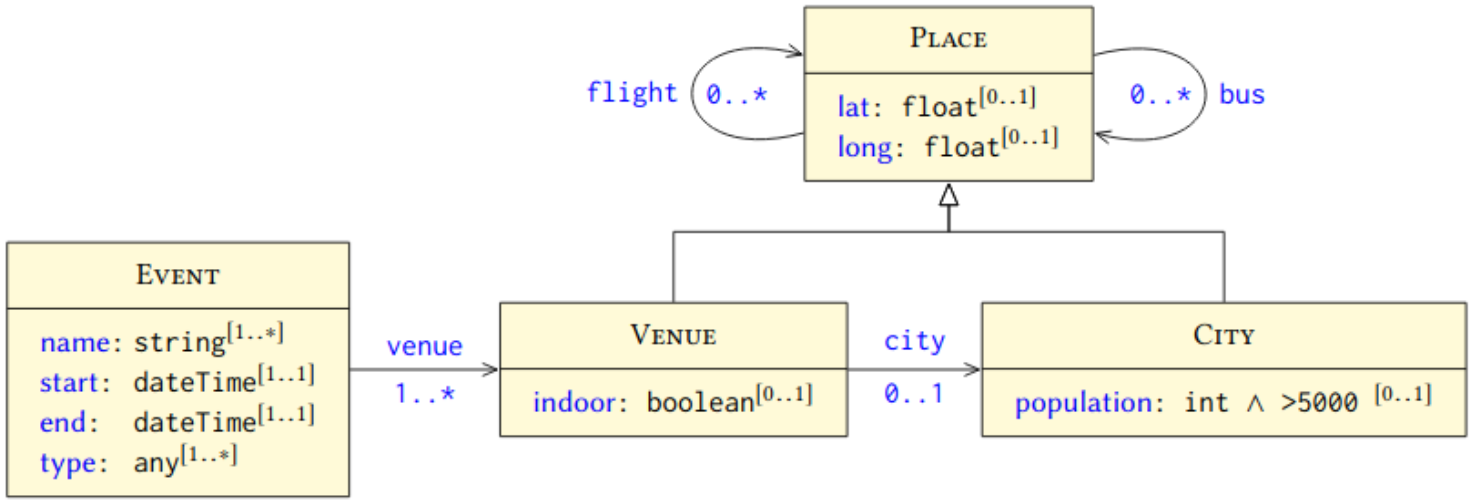
Schema - Semantic

- Defines the meaning of high-level terms
  - Classes
  - Properties (i.e., semantics of edge labels)
- RDF graphs
  - RDF Schema (RDFS)
    - Subclasses, subproperties, domain, range
  - Web Ontology Language (OWL)
    - In-depth semantics
- Defined for incomplete data graphs
  - Closed World Assumption vs Open World Assumption

Schema, Identity, Context

Schema - Validating

- Defines semantic constraints on a data graph
- Shape
  - Specify constraints on a set of nodes
  - Closed shapes vs open shapes
- Shape languages
  - Shape Expressions (ShEx)
  - Shapes Constraint Language (SHACL)



Schema, Identity, Context

Identity

- Persistent identifiers
  - Avoid naming clashes
  - e.g., IRIs in RDF graphs
- External identity links
  - Disambiguation w.r.t. external sources
  - Connected nodes represent the same entity
  - e.g., *owl:sameAs*:
- Datatypes
- Lexicalisation
  - Human-interpretable label for nodes (e.g., rdfs:label)
  - Easier to recognize real-world entities
- Existential nodes
  - Represent incomplete information (e.g., RDF's blank nodes)

Thank you

Note: The images are taken from the paper
`https://arxiv.org/pdf/2003.02320.pdf`