Analyses of RDF Triples in Sample Datasets

Jakub Stárka, **Martin Svoboda**, Irena Mlýnková

XML and Web Engineering Research Group

Charles University in Prague

Czech Republic

12 November 2012, **COLD@ISWC 2012**, Boston, MA, USA

Outline

- Introduction
  - Context
  - Objectives
  - Motivation
- Analyses
  - Characteristics
  - Experiments
- Conclusion

Context

- Querying framework
  - Data
    - Distributed and dynamic data
    - **Explicitly defined database**
      - Just datasets we want to work with…
      - … so we should know something about the data…
      - … and hopefully make the framework more effective
  - Issues
    - Indexing structures and statistics
    - Distributed query processing

Objectives

- **Data characteristics** and their…
  - **definition**
    - … to propose characteristics we want to use
  - **detection**
    - … to describe data we want to work with
  - **publishing**
    - … to publish characteristics to the others
  - **exploitation**
    - … to make data processing more effective

Motivation

- **Optional** scenarios
  - Knowledge of characteristics may help us…
    - … to propose more efficient indices in general or to adjust their usage depending on situations
    - …
- **Required** scenarios
  - Knowledge of characteristics is required…
    - … even to create instances of indices because they require parameterization
    - …

Characteristics

- Groups of proposed characteristics
  - **Terms**
    - Features of URI references and literals
  - **Triples**
    - Features of components of RDF triples
  - **Graphs**
    - More complex features of sets of triples

Experiments

- **Sample datasets**
  - ACM Publications
  - Czech DBPedia
  - English DBPedia
  - Gene Ontology
  - Movie Database

Terms

- Motivation
  - **Terms** (or at least their substrings) **often repeat**
- Questions
  - What is the average length of terms?
    - URI references
    - Prefixes of URLs before optional fragment parts
    - Literals

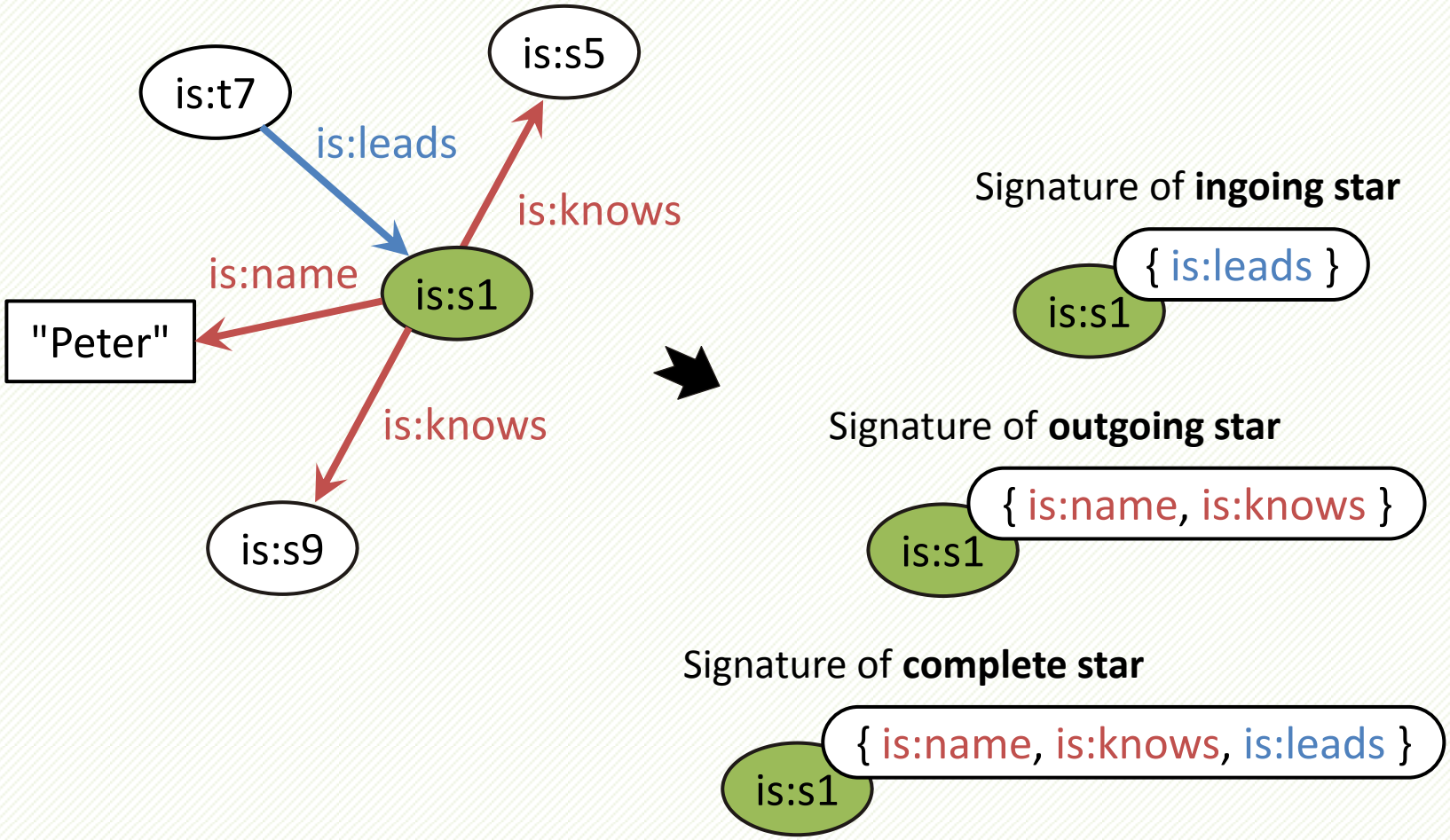
Projections

- Motivation
  - **There are often more triples** with the same…
    - subjects / predicates / objects
    - pairs of subject and predicate / …
- Definitions
  - **Projection** on a given component/s for a particular term value/s **as a set of corresponding triples**
  - **Projections are divided into classes** by their sizes

Projections

- Questions
  - **How frequent are projections with size…**
    - … equal to one?
    - … greater than one?

Stars


is:t7
is:s5
is:leads
is:knows
is:name
is:s1
"Peter"
is:knows
is:s9
Signature of **ingoing star**
{ is:leads }
is:s1
Signature of **outgoing star**
{ is:name, is:knows }
is:s1
Signature of **complete star**
{ is:name, is:knows, is:leads }
is:s1

Stars

- Definitions
  - **Signature of** an ingoing/outgoing/complete **star** around a given vertex **as a set of predicates of** ingoing/outgoing/all **edges** to/from this vertex
  - **Vertices are divided into classes** using signatures
- Motivation
  - SPARQL queries often contain star graph patterns
- Questions
  - **Which star signatures should be indexed?**

Paths

- Definitions
  - **Signature of a** (directed and disjoint) **path as a sequence of predicates of its edges**
  - **Paths are divided into classes** using signatures
- Questions
  - **Paths of which lengths** should be considered?
  - **Which path signatures should be indexed?**

Conclusion

- **Characteristics**
  - Terms – lengths of terms and prefixes
  - Triples – projections
  - Graphs – stars and paths
- **Exploitation**
  - Storing, indexing and querying of RDF data

Thank you for your attention…

XML and Web Engineering Research Group

Charles University in Prague