软件新技术与产业化协同创新中心
Collaborative Innovation Center of Novel Software Technology and Industrialization

Entity Linkage in the Linked Data:
approaches and analysis

Wei Hu (whu@nju.edu.cn)

Department of Computer Science and Technology

National Key Laboratory for Novel Software Technology

Nanjing University, China

Université d'Artois, 2016.6.7

Websoft (http://ws.nju.edu.cn)

Yuzhong Qu

Wei Hu

Gong Cheng

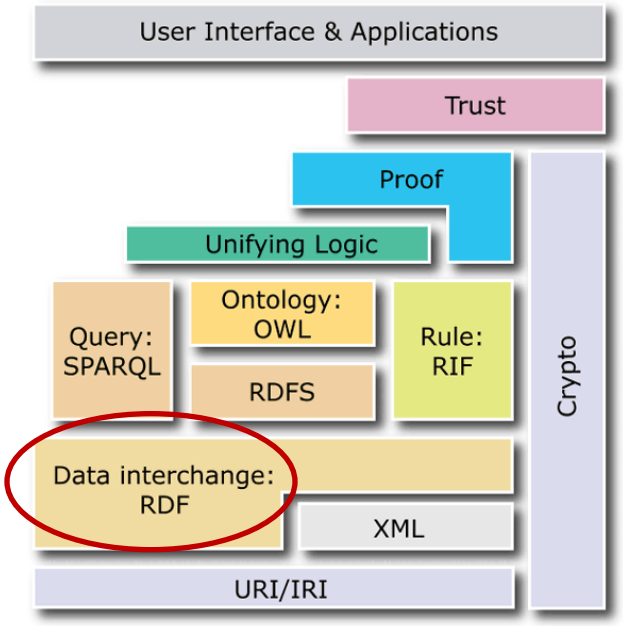
- Research topics
  - **Semantic Web**
  - **Web science**
  - **Big data**
- Academic records
  - Papers
    - WWW, IJCAI, AAAI, ISWC …
    - Best paper award & nominee
  - Grants: 863, NSFC …
- Collaborations
  - Stanford, VUA, KIT, Aberdeen …
  - IBM, Samsung, ZTE …

Outline

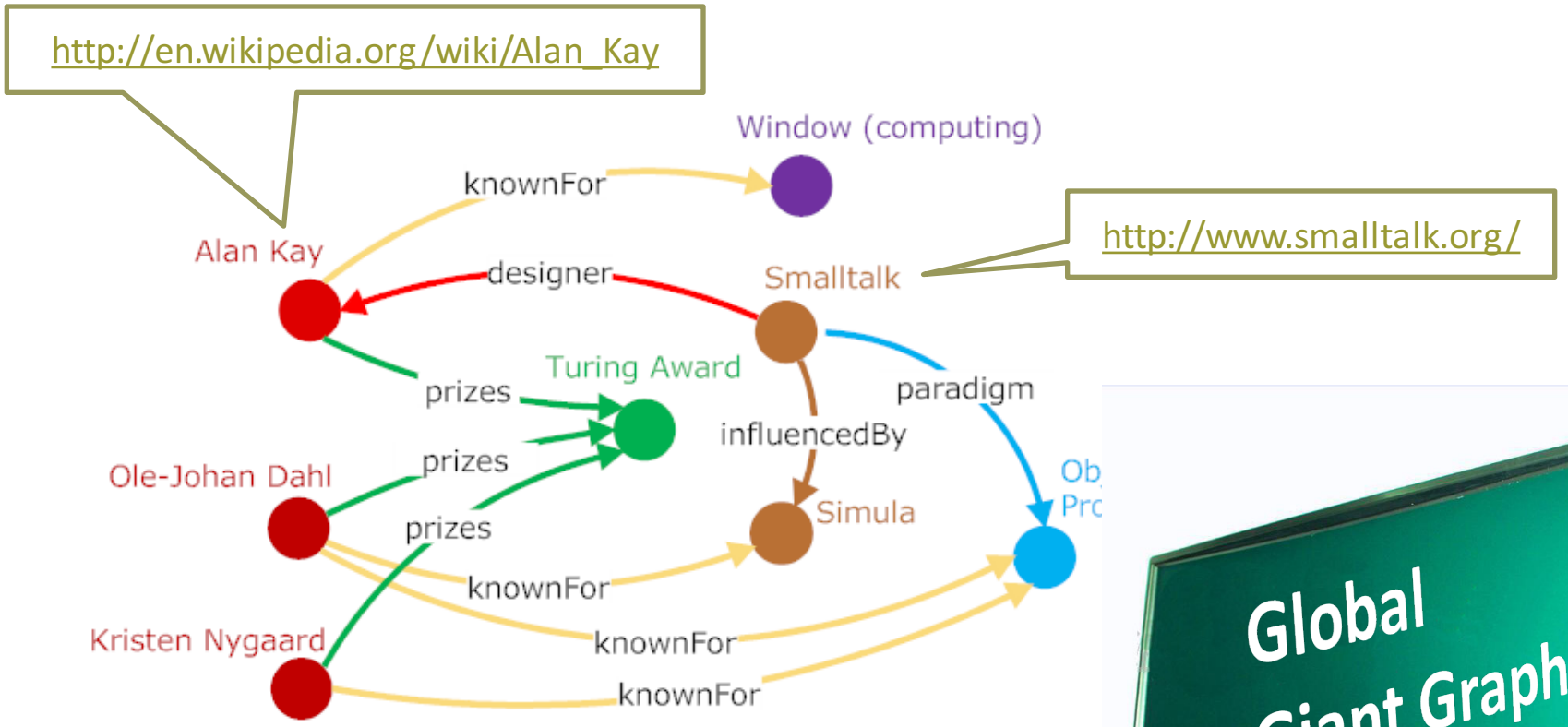
- **Introduction to Semantic Web and entity linkage**
- A bootstrapping approach to entity linkage
- Link analysis of biomedical linked data
- (Two applications)

Semantic Web

- Semantic Web was a thought from Tim Berners-Lee
- Give formal meanings to Web information – **semantics**
  - Web 1.0 (page) → Web 2.0 (social) → Web 3.0 (**a web of data**)
- Semantic Web is about
  1. **common formats** for
     - integration and combination of data drawn from diverse sources
  2. **languages** for
     - recording how the data relates to real-world objects



RDF (Resource Description Framework)



RDF triple: ***< subject, predicate, object >***

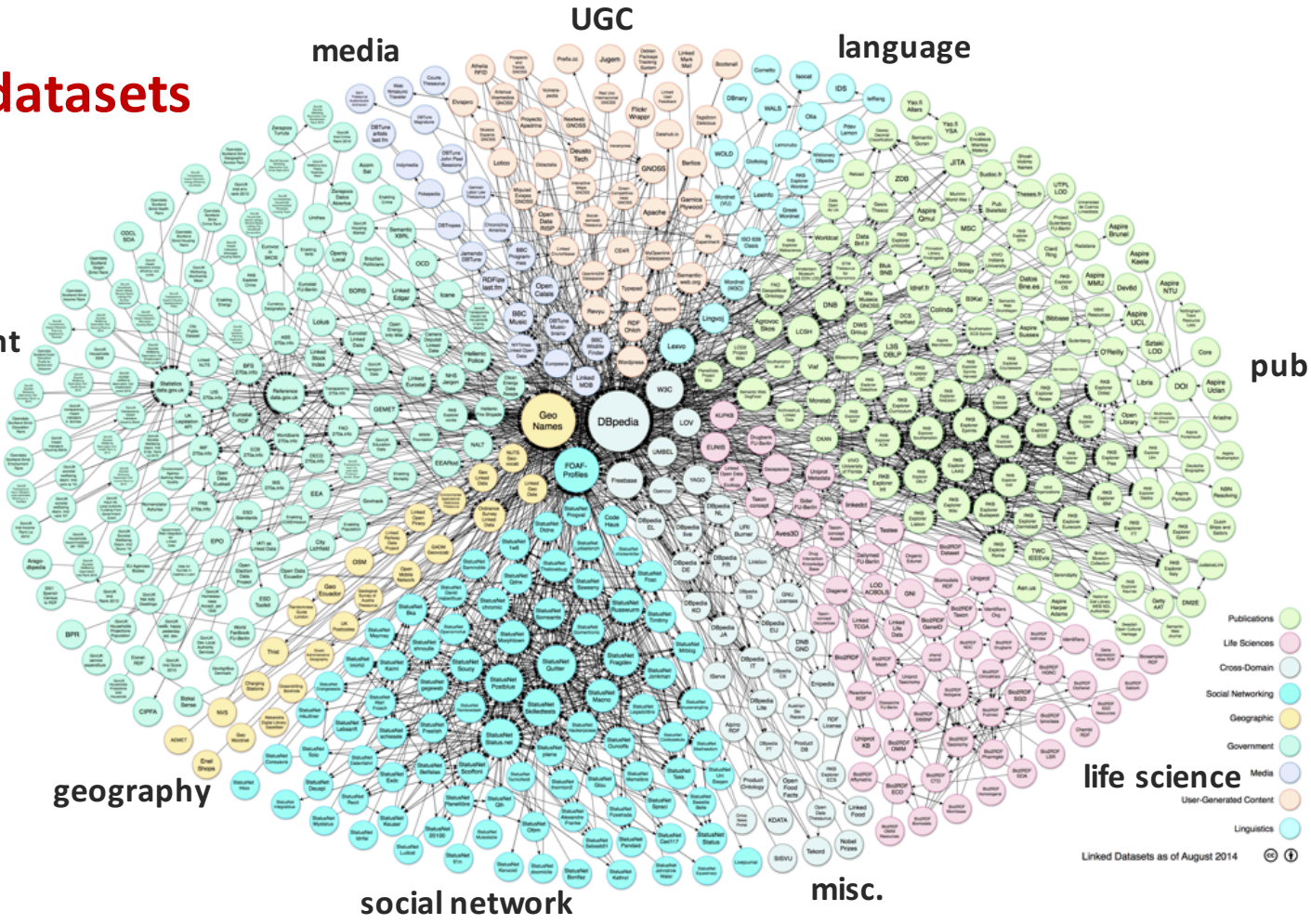
Linked data

- As a realization of Semantic Web
  - **Linked Data** refers to a collection of interrelated datasets
    - Used for **large-scale integration** of, **reasoning** on, data on the Web

- **Linked data principles**
  1. Use **URIs** to name things
  2. Use **HTTP URIs** (can be "dereferenced")
  3. Provide useful information using the open Web standards (e.g. **RDF**)
  4. Include **links** to other related things

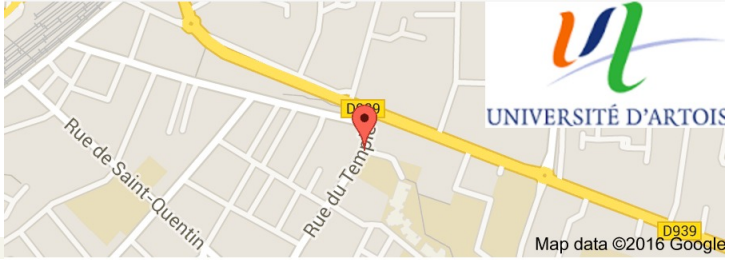
Linking open data (LOD) cloud

1,014 datasets



Knowledge graph

- Knowledge Graph is a **knowledge base** used by Google to enhance its search engine's search results with semantic search information gathered from a wide variety of sources
  - Nodes: entities or concepts
  - Edges: attributes or relations



Artois University ★ Website Directions

University in Arras, France

Artois University is a French university, based in Arras. It is under the umbrella of the Academy of Lille and is a member of the European Doctoral College Lille-Nord-Pas de Calais. Wikipedia

Address: 9 Rue du Temple, 62000 Arras

Enrollment: 10,956 (2013)

Phone: 03 21 60 37 00

Founded: 1992

Academic staff: 846

Suggest an edit · Own this business?

Send to your phone Send

Profiles

Facebook Twitter YouTube

Reviews Write a review Add a photo

8 Google reviews

People also search for View 5+ more

Lille University of Scienc... Villeneuve-...

University of Valencie... Valenciennes

Lille 2 University of Health... Lille

University of Lille Nord de F... Villeneuve-...

Charles de Gaulle University... Villeneuve-...

Entity linkage

- Semantic Web data reach a scale in billions of entities
- Many different entities refer to the **same** real-world thing
  - Typically denoted by URIs, from distributed data sources
    - e.g. Wei Hu
      - http://data.semanticweb.org/person/wei-hu
      - http://ws.nju.edu.cn/people/whu
      - http://ontoworld.org/wiki/Special:URIResolver/Wei_Hu
      - …
- **Entity linkage:** link different entities that refer to the same object
  - a.k.a. coreference resolution, entity matching …
  - Out of 31B RDF statements, less than 500M are links across sources

Outline

- Introduction to Semantic Web and entity linkage
- **A bootstrapping approach to entity linkage**
- Link analysis of biomedical linked data
- (Two applications)

Background

- In LOD, millions of entities have already been linked
  - However, potential candidates are still **numerous**
- Current solutions
  1. **Equivalence reasoning**
     - ✓ owl:sameAs, inverse functional properties …
     - ✓ **At present, probably miss many potential candidates**
  2. **Similarity computation** (also in the database area)
     - ✓ Compare properties and values of entities
     - ◆ **Inaccurate (heterogeneity), less scalable (pairwise comparison)**
  3. To improve, **machine learning**
     - ◆ **Time-consuming, labor-intensive to build a large-scale training set**

Definition

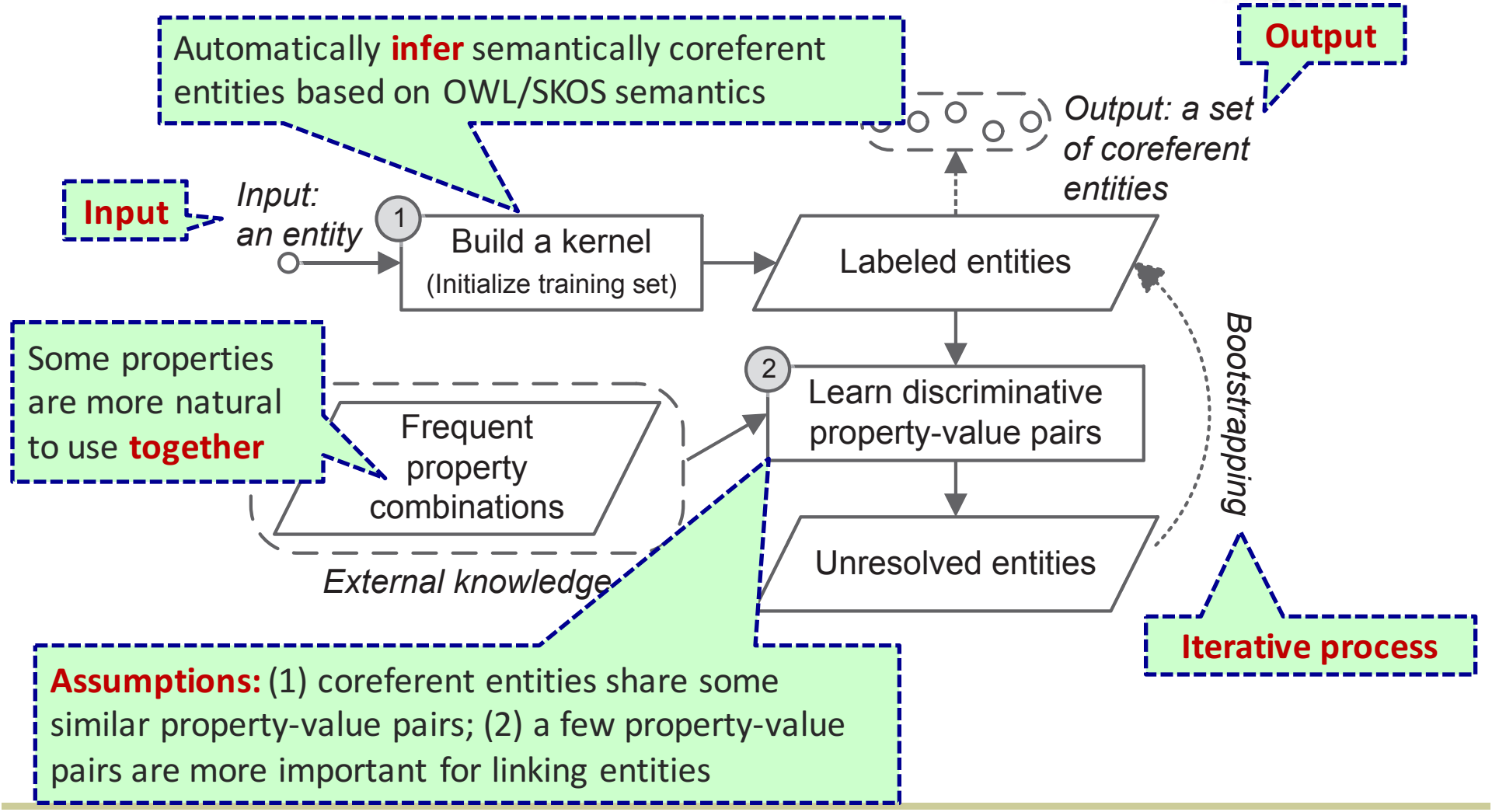
How to combine? Our solution: bootstrapping

- **Query-driven** entity linkage

> **Definition 1.** Let **U** be the set of entities in a set **D** of data sources. Given an entity $u \in \mathbf{U}$, the entity linkage for *u* is to query a subset $\mathbf{E}(u) \subseteq \mathbf{U}$ of entities for which a relation *ε* holds:
>
> $$\mathbf{E}(u) = \{v \in \mathbf{U} \mid (u, v) \in \mathcal{E}\}$$
>
> where *ε* links all the entities in **U** that refer to the same object as *u* does, i.e. are **coreferent** with *u*.

- Use scenarios
  1. Search / browsing – a system knows "what to link" only at query time
  2. Analyze small portions of a very large dataset to answer on-demand queries

Our contribution



Running example

dbpedia:Nanjing (DBpedia)	rdfs:label owl:sameAs	“ **Nanjing** ” ***geo:1799962***
geo:1799962 (GeoNames)	geo:lat geo:long geo:alternateName	“ 32 N ” “ 118 E ” “ **Nanjing** ” “ Nan-ching ”
fb:m.05gqy (Freebase)	rdfs:label geo:lat geo:long	“ **Nanjing** ” “ 32 N ” “ 118 E ”
ex:NationalCity	geo:long geo:lat	“ 117 W ” “ 32 N ”

Experiment

- Dataset
  - Billion Triples Challenge (BTC) 2011
- Testing entities
  - **Top-50** in 364 thousand query logs

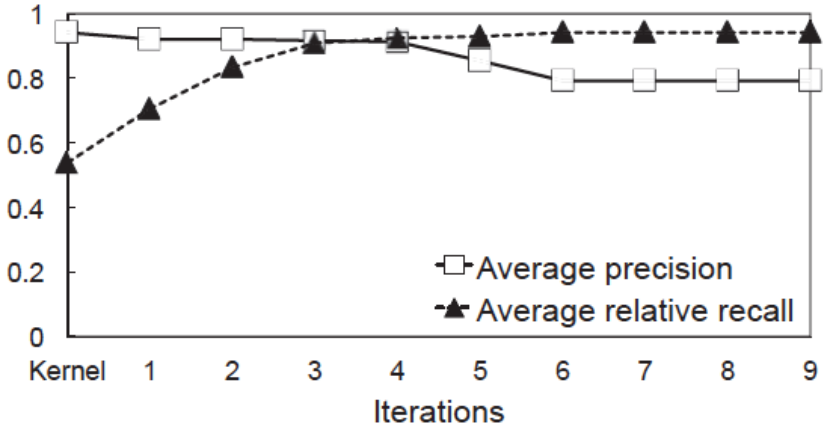
People	15	Places	10
Tech terms	8	Music / movies	5
Universities	4	Companies	3
Publications	2	others	3

Entities	> 100 million
RDF stat.	> 2 billlion
Same-as stat.	3,446,029
IFP stat.	1,799,976
FP stat.	2,279,474
Exact-match stat.	22,398
Cardinality stat.	148
Has-key stat.	2
Different-from stat.	691
All-different stat.	89

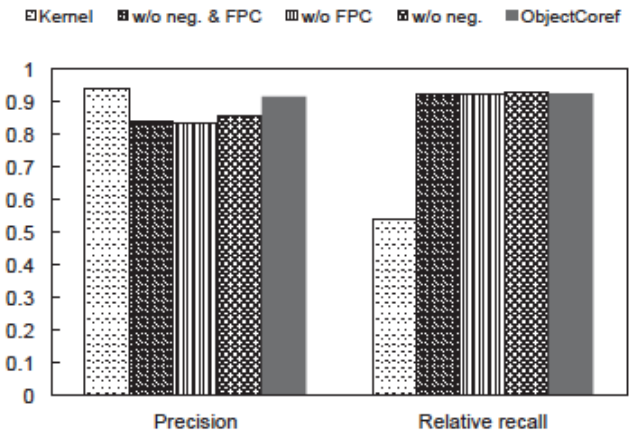
- Evaluation procedure and metrics
  - 30 graduates, 2 judges + 1 arbitrator / link, **Fleiss's $\kappa = 0.8$** (sufficient agree)
  - Precision & **relative recall** (RR)
    - RR = correct links in one system / total correct unique links in all systems

Experiment

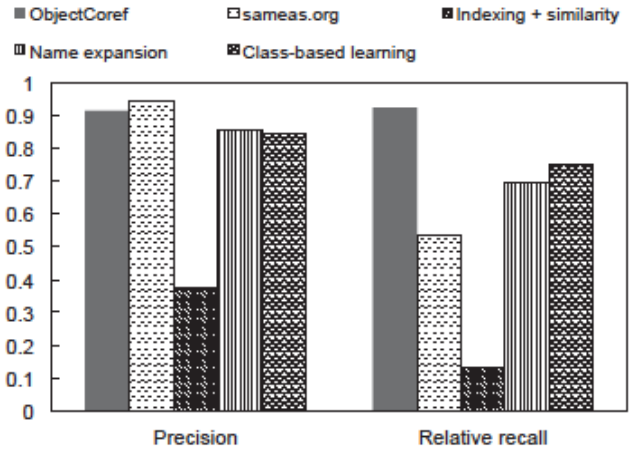
- Bootstrapping curve
  - Maximum iteration = 4
  - Discriminability threshold = 0.05



- Linkage accuracy



(a) Different components



(b) ObjectCoref vs. others

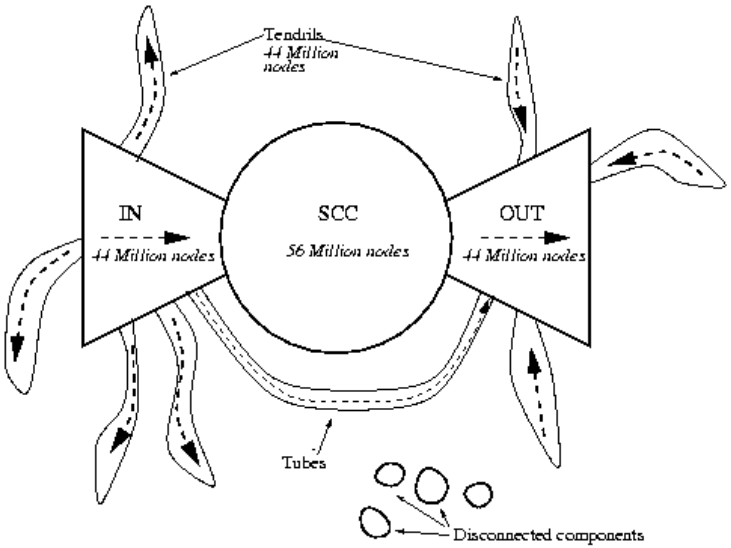
- Running time on 5,000 samples: **avg. 11.3 links in 12.6s**

Outline

- Introduction to Semantic Web and entity linkage
- A bootstrapping approach to entity linkage
- **Link analysis of biomedical linked data**
- (Two applications)

Background

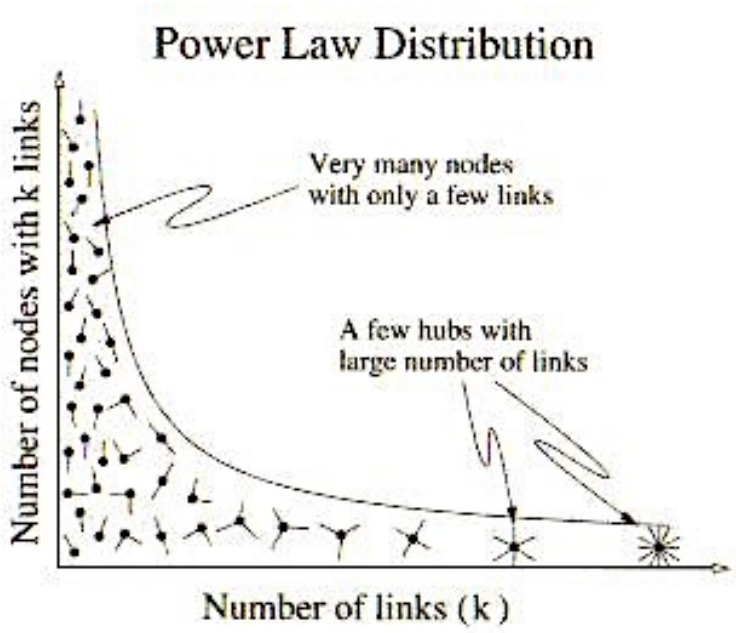
- **Network analysis** has long been used to study link structures
  - Network medicine: cellular networks and implications
  - The “bow tie” structure of the Web
- **Linked data for the life sciences**
  - e.g. Bio2RDF, Chem2Bio2RDF, Neurocommons, W3C LODD
    - Millions of links over hundreds of datasets in overlap
  - Network analysis can help
    - understand structures to express data
    - facilitate large-scale data integration
    - improve overall quality of biomedical data



No such analysis yet!

Preliminaries

- **Graph**: nodes and edges
  - (outgoing / incoming) degree
    - Sink, source, isolated node



- **Power law** distribution
  - $p(x) \propto x^{-\alpha}$
  - **Scale-free**
- Weakly connected component
  - **Size**: number of nodes
- Average distance
- Clustering coefficient

→ **Small-world** phenomenon

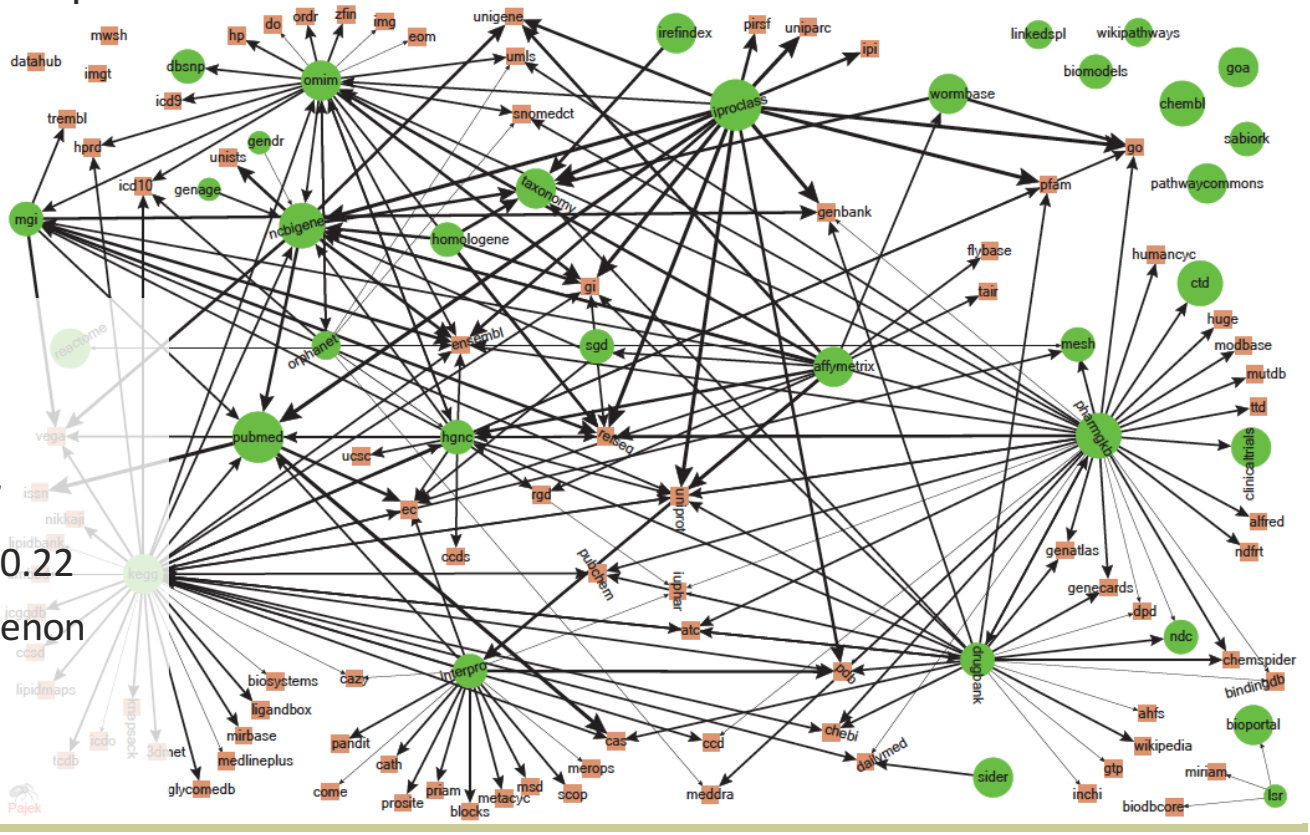
Our contribution

- We conduct an empirical link analysis of Bio2RDF
  - Bio2RDF is an open source project that uses Semantic Web technologies to build and provide the largest network of life science Linked Data
    - Ensure the significance of our empirical study
  1. Dataset link analysis (using RDF data model)
  2. **Entity link analysis (using a special kind of cross-references)**
  3. Term link analysis (using ontology matching)
     - For each perspective, we investigate the graph features of Bio2RDF vis-à-vis what has been previously reported
     - Symmetry and transitivity of entity links
     - Benchmark to evaluate entity matching approaches

Dataset overview

What is the status of Bio2RDF?

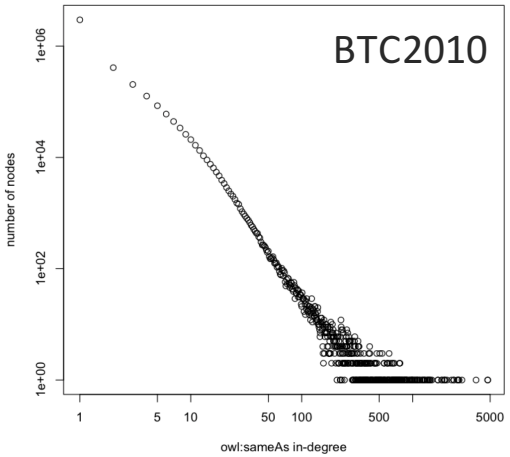
- 35 datasets, 11B RDF triples
- 1B entities
- 2K classes
- 4K properties



Observations

1. Well linked
2. Average distance = 2.77
   Clustering coefficient = 0.22
   → small-world phenomenon
3. Hubs and authorities
4. Good resilience

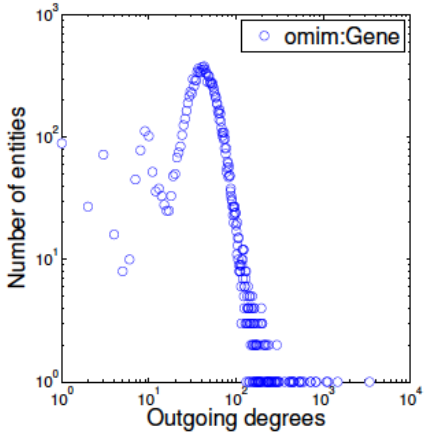
Entity link analysis

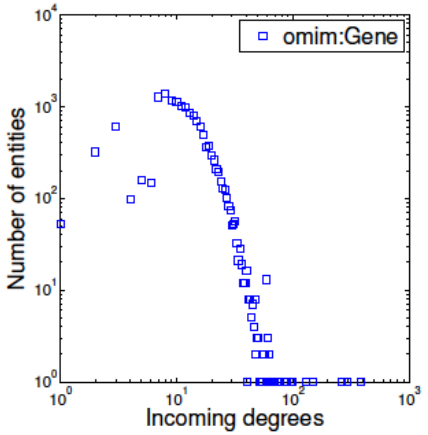


How well do entities link to each other?

- 76% entity links from a special kind of RDF triples
  - e.g. <kegg:D03455, kegg:x-drugbank, drugbank:DB00002>
  - x-relations have **under-specified semantics**
    - Refer to a related resource, e.g. article
    - Truly identical

- Degree distribution
  - Three types of entities in OMIM, NCBI, KEGG
  - **Do not follow power law**
    - Exponent is too large (close to 5), p-values is too small (close to 0)

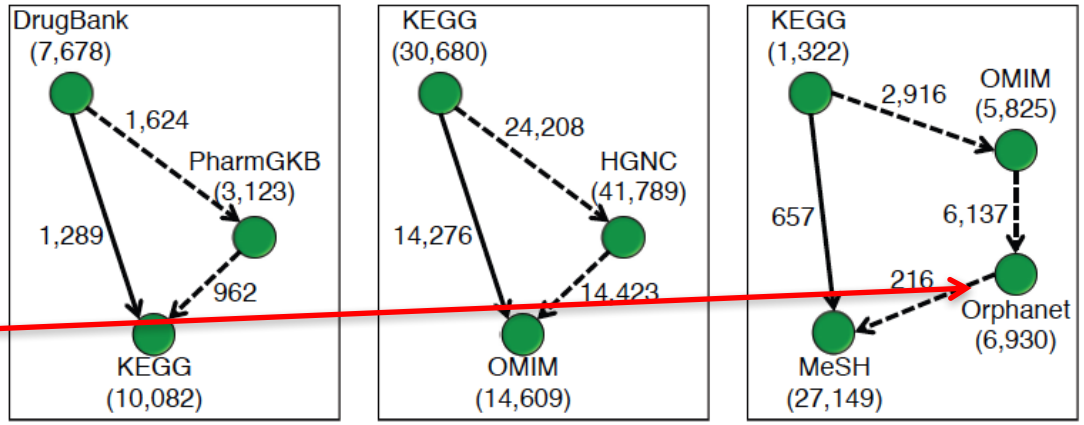




Symmetry and transitivity

- **Symmetry**
  - Borrow and reverse
  - Different classes
    - Phenotype vs. Disorder
- **Transitivity**
  - Weak intermediate
  - Modeling divergence
    - Even hard to human

	Forward	Backward	Reciprocal	Malposed	Missing	Total
DrugBank—KEGG	1,289	2,155	1,964	485	995	3,444
DrugBank—PharmGKB	1,624	1,619	3,210	4	29	3,243
OMIM—HGNC	14,274	14,423	28,514	6	177	28,697
OMIM—Orphanet	6,137	2,600	4,464	2,523	1,750	8,737



(a) Drugs (b) Human genes (c) Diseases

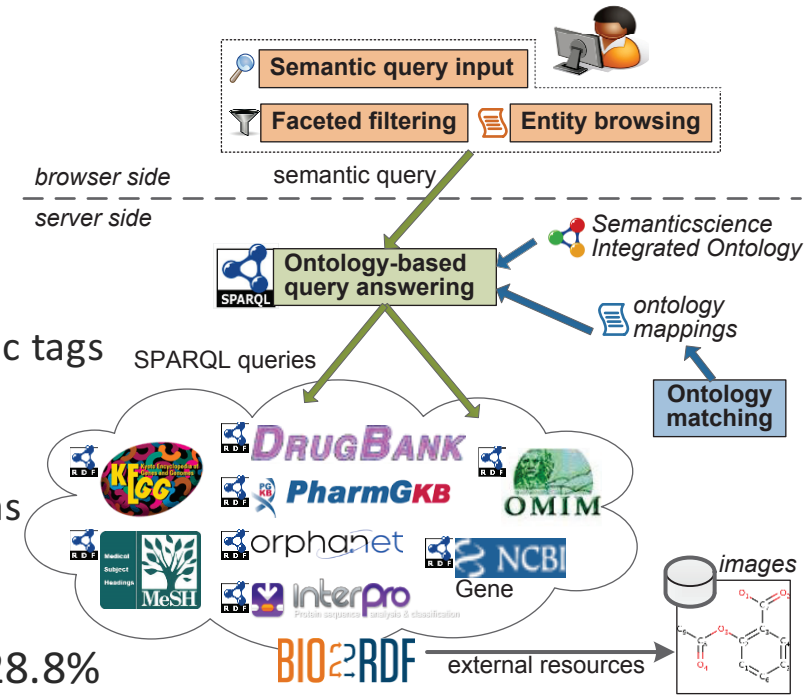
	Direct links	Transitive paths	Identical ending entities	Different ending entities	Missing direct	Missing transitive	Total
Drugs	1,289	954	946	6	2	343	1,297
Human genes	14,276	14,250	14,236	5	9	40	14,290
Diseases	657	33	8	18	7	649	682

Discussion of findings

- Entity link graph **does not share** the same characteristics with the Hypertext / Semantic Web
  - Degree distribution does not follow power law
- A **dominated** part of entities have been linked using x-relations, but their intended semantics **differs**
  - Classes are identical or equivalent → entity links represent logical equivalence
- Symmetric and transitive entity links exist, but their effectiveness is **weakened** due to the small number
  - Meanings of entity links may shift during transitive
  - KEGG, DrugBank and OMIM are the most prominent knowledge bases

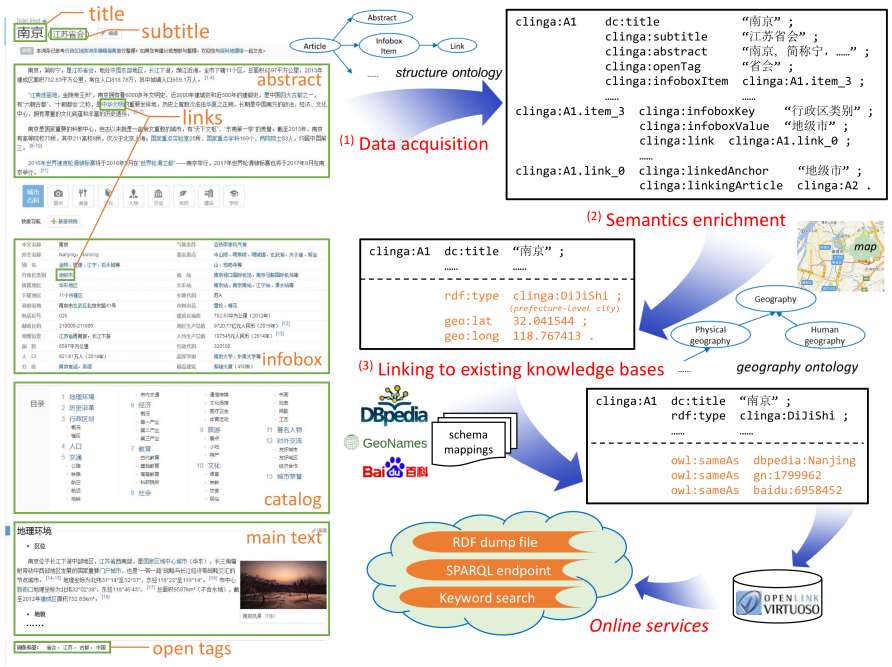
Applications: BioSearch

- **Keyword search** is the most popular paradigm for information retrieval
  - Keywords can be ambiguous and have multiple meanings
    - **Semantic search** aims to improve search accuracy by understanding user intent and search context
  - **Heterogeneity** between local schemas
- Our solution
  1. Semantic query + faceted filtering
     - Not only plain keywords but also semantic tags
  2. Onto-based query answering
     - Rewrite queries from SIO to local schemas
  3. Entity browsing
  - Result: effectiveness +22.4%, usability +28.8%



Applications: Clinga

- Chinese geographical data is small scale, e.g. 4.6% in GeoNames
- **C**hinese **lin**ked **g**eographical d**a**taset (Clinga)
  1. Extract data from the largest Chinese wiki encyclopedia
  2. Design a geo-ontology to classify geographical entity types
  3. Automatic discovery of links to existing knowledge bases
     - Result: 624K entities, 230K links
- Use scenario
  - Major knowledge base for answering Chinese geographical questions in our National Higher Education Entrance Examination (called *GaoKao*)



Conclusion

- Entity linkage is to link different entities that refer to the same real world object
- Large scale and heterogeneity are challenging existing entity linkage solutions
- Entity linkage approaches often involve knowledge representation, data mining, network analysis, crowdsourcing and many other techniques

Thank you for your invitation and time!

Comments?

Contact: Wei Hu (whu@nju.edu.cn)