KNOWLEDGE GRAPH IDENTIFICATION

Jay Pujara
jay@cs.umd.edu

11/5/2014
presented at Carnegie Mellon University
Motivating Problem: Opportunities

Internet

Massive source of publicly available information

Extraction

Cutting-edge IE methods

Knowledge Graph (KG)

Structured representation of entities, their labels and the relationships between them
Knowledge Graphs in the wild
Motivating Problem: Real Challenges

Internet

Noisy!

Extraction

Difficult!

Knowledge Graph

Contains many errors and inconsistencies
Examples of NELL errors
Kyrgyzstan has many variants:

- Kyrgystan
- Kyrgistan
- Kyrghyzstan
- Kyrgzstan
- Kyrgyz Republic

Saudi Cultural Days in the Kyrgyz Republic has concluded its activities in the capital Bishkek in the weekend in a special ceremony held on this occasion. The event was attended by Deputy Minister of Culture and Tourism of the Kyrgyz Republic Koulev Mirza; Kyrgyzstan’s Ambassador to Saudi Arabia Jusupbek Sharipov; the Saudi Embassy Acting Chargé d'affaires to Kyrgyzstan, Mari bin Barakah Al-Derbas and members of the embassy staff, in the presence of a heavy turnout of Kyrgyz citizens.

The Days of Culture of Saudi Arabia in Kyrgyzstan will be held from 6 to 9 May.

Refugees are often from areas where conflict is historically embedded and marked in ideology and injustice. The Tsarnaev family emigrated from the Chechen diaspora in Kyrgyzstan, a region Stalin deported the Chechens to in 1943. After the fall of the Berlin Wall in 1991, Chechens engaged in a battle for independence from Russia that led to the Tsarnaevs' petition for refugee status in the early
Kyrgyzstan is labeled a bird and a country.

Erik Kleyheeg has just returned from Lesvos with some new bird images. Included here are: Common Scops-Owl, Wood Warbler, Spanish Sparrow, Red-throated Pipit, Eurasian Chiff-chaff, and Cretzschmar’s Bunting.

Anssi Kullberg has sent along some great trip reports to unusual places, including Kyrgyzstan, Pakistan.

Kyrgyzstan (IPA: [kʊrgʲɪˈstan]; Russian: Кыргызстан) is a country located in Central Asia. Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Its capital and largest city is Bishkek.
Kyrgyzstan’s location is ambiguous – Kazakhstan, Russia and US are included in possible locations

Kyrgyzstan U.S. Air Base Future Unclear

A Central Asian country of incredible natural beauty and proud nomadic traditions, most of Kyrgyzstan was formally annexed to Russia in 1876. The Kyrgyz staged a major revolt against the Tsarist Empire in 1916 in which almost one-sixth of the Kyrgyz population was killed. Kyrgyzstan became a Soviet republic in 1936 and
Violations of ontological knowledge

- Equivalence of co-referent entities (sameAs)
  - SameEntity(Kyrgyzstan, Kyrgyz Republic)
- Mutual exclusion (disjointWith) of labels
  - MUT(bird, country)
- Selectional preferences (domain/range) of relations
  - RNG(countryLocation, continent)

Enforcing these constraints requires jointly considering multiple extractions across documents
Motivating Problem (revised)

Internet

(noisy) Extraction Graph

Large-scale IE

Knowledge Graph

Joint Reasoning

(Pujara et al., ISWC13)
Transformation

Input Graph

Available but inappropriate for analysis

Graph Identification

Output Graph

Appropriate for further analysis
Motivation: Different Networks

**Communication Network**
- Nodes: Email Address
- Edges: Communication
- Node Attributes: Words

**Organizational Network**
- Nodes: Person
- Edges: Manages
- Node Labels: Title
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

Label: CEO Manager Assistant Programmer
Graph Identification

Input Graph: Email Communication Network
Output Graph: Social Network

What’s involved?
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

• What’s involved?
  • Entity Resolution (ER): Map input graph nodes to output graph nodes
Graph Identification

- What’s involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes
  - Link Prediction (LP): Predict existence of edges in output graph

Input Graph: Email Communication Network
Output Graph: Social Network
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

What’s involved?

- **Entity Resolution (ER):** Map input graph nodes to output graph nodes
- **Link Prediction (LP):** Predict existence of edges in output graph
- **Node Labeling (NL):** Infer the labels of nodes in the output graph
Most work looks at these tasks in isolation.

In graph identification, they are:

- Evidence-Dependent – Inference depend on observed input graph e.g., ER depends on input graph
- Intra-Dependent – Inference within tasks are dependent e.g., NL prediction depend on other NL predictions
- Inter-Dependent – Inference across tasks are dependent e.g., LP depend on ER and NL predictions
KNOWLEDGE
GRAPH
IDENTIFICATION

Pujara, Miao, Getoor, Cohen, ISWC 2013 (best student paper)
Motivating Problem (revised)

(From Pujara et al., ISWC13)
**Knowledge Graph Identification**

**Problem:**

Extraction Graph \[ \rightarrow \] Knowledge Graph Identification

**Solution:** *Knowledge Graph Identification* (KGI)

- Performs *graph identification*:
  - entity resolution
  - node labeling
  - link prediction
- Enforces *ontological constraints*
- Incorporates *multiple uncertain sources*
Illustration of KGI: Extractions

**Uncertain Extractions:**
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)
**Illustration of KGI: Ontology + ER**

**Uncertain Extractions:**
- 5: Lbl(Kyrgyzstan, bird)
- 7: Lbl(Kyrgyzstan, country)
- 9: Lbl(Kyrgyz Republic, country)
- 8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

**Extraction Graph**

(Pujara et al., ISWC13)
Illustration of KGI: Ontology + ER

**Uncertain Extractions:**
- 0.5: Lbl(Kyrgyzstan, bird)
- 0.7: Lbl(Kyrgyzstan, country)
- 0.9: Lbl(Kyrgyz Republic, country)
- 0.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

**Ontology:**
- Dom(hasCapital, country)
- Mut(country, bird)

**Entity Resolution:**
- SameEnt(Kyrgyz Republic, Kyrgyzstan)

(Annotated) Extraction Graph

(Pujara et al., ISWC13)
Illustration of KGI

Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)  
.7: Lbl(Kyrgyzstan, country)  
.9: Lbl(Kyrgyz Republic, country)  
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

Ontology:
Dom(hasCapital, country)  
Mut(country, bird)

Entity Resolution:
SameEnt(Kyrgyz Republic, Kyrgyzstan)

(Annotated) Extraction Graph

After Knowledge Graph Identification
Modeling Knowledge
Graph Identification
Viewing KGI as a probabilistic graphical model

(Pujara et al., ISWC13)
Background: Probabilistic Soft Logic (PSL)

(Broechele et al., UAI10; Kimming et al., NIPS-ProbProg12)

- Templating language for hinge-loss MRFs, very scalable!
- Model specified as a collection of logical formulas

\[
\text{SAMEENT}(E_1, E_2) \sim \text{LBL}(E_1, L) \Rightarrow \text{LBL}(E_2, L)
\]

- Uses soft-logic formulation
  - Truth values of atoms relaxed to $[0,1]$ interval
  - Truth values of formulas derived from Lukasiewicz t-norm
Background: PSL Rules to Distributions

- Rules are *grounded* by substituting literals into formulas

\[
w_{EL}: \text{SAMEENT(Kyrgyzstan, Kyrgyz Republic)} \sim \text{LBL(Kyrgyzstan, country)} \Rightarrow \text{LBL(Kyrgyz Republic, country)}
\]

- Each ground rule has a **weighted distance to satisfaction** derived from the formula’s truth value

\[
P(G \mid E) = \frac{1}{\mathcal{Z}} \exp \left[ - \sum_{r \in R} w_r \phi_r(G) \right]
\]

- The PSL program can be interpreted as a joint probability distribution over all variables in knowledge graph, conditioned on the extractions

(Pujara et al., ISWC13)
Background: Finding the best knowledge graph

- MPE inference solves $\max_G P(G)$ to find the best KG

- In PSL, inference solved by convex optimization

- Efficient: running time empirically scales with $O(|R|)$
  (Bach et al., NIPS12)
PSL Rules for KGI Model
PSL Rules: Uncertain Extractions

\[ w_{CR-T} : \text{CANDREL}_T(E_1, E_2, R) \]
\[ w_{CL-T} : \text{CANDLBL}_T(E, L) \]

Predicate representing uncertain relation extraction from extractor T

Relation in Knowledge Graph

\[ \Rightarrow \text{REL}(E_1, E_2, R) \]
\[ \Rightarrow \text{LBL}(E, L) \]
PSL Rules: Entity Resolution

\[ w_{EL} : SameEnt(E_1, E_2) \land Lbl(E_1, L) \Rightarrow Lbl(E_2, L) \]

\[ w_{ER} : SameEnt(E_1, E_2) \land Rel(E_1, E, R) \Rightarrow Rel(E_2, E, R) \]

\[ w_{ER} : SameEnt(E_1, E_2) \land Rel(E, E_1, R) \Rightarrow Rel(E, E_2, R) \]

- SameEnt predicate captures confidence that entities are co-referent.
- Rules require co-referent entities to have the same labels and relations.
- Creates an equivalence class of co-referent entities.

(Pujara et al., ISWC13)
PSL Rules: Ontology

Inverse:

\[ w_0 : \text{INV}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_2, E_1, S) \]

Selectional Preference:

\[ w_0 : \text{DOM}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_1, L) \]
\[ w_0 : \text{RNG}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_2, L) \]

Subsumption:

\[ w_0 : \text{SUB}(L, P) \quad \land \quad \text{LBL}(E, L) \quad \Rightarrow \quad \text{LBL}(E, P) \]
\[ w_0 : \text{RSUB}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_1, E_2, S) \]

Mutual Exclusion:

\[ w_0 : \text{MUT}(L_1, L_2) \quad \land \quad \text{LBL}(E, L_1) \quad \Rightarrow \quad \neg \text{LBL}(E, L_2) \]
\[ w_0 : \text{RMUT}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \neg \text{REL}(E_1, E_2, S) \]

Adapted from Jiang et al., ICDM 2012
[\phi_1] \text{CANDLBL}_{struct}(Kyrgyzstan, bird) \\
\Rightarrow \text{LBL}(Kyrgyzstan, bird)

[\phi_2] \text{CANDREL}_{pat}(Kyrgyz Rep., Asia, locatedIn) \\
\Rightarrow \text{REL}(Kyrgyz Rep., Asia, locatedIn)

[\phi_3] \text{SAMEENT}(Kyrgyz Rep., Kyrgyzstan) \\
\land \text{LBL}(Kyrgyz Rep., country) \\
\Rightarrow \text{LBL}(Kyrgyzstan, country)

[\phi_4] \text{DOM}(locatedIn, country) \\
\land \text{REL}(Kyrgyz Rep., Asia, locatedIn) \\
\Rightarrow \text{LBL}(Kyrgyz Rep., country)

[\phi_5] \text{MUT}(country, bird) \\
\land \text{LBL}(Kyrgyzstan, country) \\
\Rightarrow \neg \text{LBL}(Kyrgyzstan, bird)
Probability Distribution over KGs

\[ P(G \mid E) = \frac{1}{Z} \exp \left[ - \sum_{r \in R} w_r \varphi_r (G, E) \right] \]

- \text{CandLbl}_T(\text{kyrgyzstan}, \text{bird}) \Rightarrow \text{Lbl}(\text{kyrgyzstan}, \text{bird})
- \text{Mut}(\text{bird}, \text{country}) \Rightarrow \neg\text{Lbl}(\text{kyrgyzstan}, \text{bird})
- \text{SameEnt}(\text{kyrgz republic}, \text{kyrgyzstan}) \neg\text{Lbl}(\text{kyrgz republic}, \text{country}) \Rightarrow \text{Lbl}(\text{kyrgyzstan}, \text{country})
Evaluation
# Two Evaluation Datasets

<table>
<thead>
<tr>
<th></th>
<th>LinkedBrainz</th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Community-supplied data about musical artists, labels, and creative works</td>
<td>Real-world IE system extracting general facts from the WWW</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>Realistic synthetic noise</td>
<td>Imperfect extractors and ambiguous web pages</td>
</tr>
<tr>
<td><strong>Candidate Facts</strong></td>
<td>810K</td>
<td>1.3M</td>
</tr>
<tr>
<td><strong>Unique Labels and Relations</strong></td>
<td>27</td>
<td>456</td>
</tr>
<tr>
<td><strong>Ontological Constraints</strong></td>
<td>49</td>
<td>67.9K</td>
</tr>
</tbody>
</table>

*(Pujara et al., ISWC13)*
## NELL Evaluation: two settings

<table>
<thead>
<tr>
<th>Target Set: restrict to a subset of KG</th>
<th>Complete: Infer full knowledge graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Jiang, ICDM12)</td>
<td></td>
</tr>
</tbody>
</table>

- **Closed-world model**
- Uses a target set: subset of KG
- Derived from 2-hop neighborhood
- Excludes trivially satisfied variables

- **Open-world model**
- All possible entities, relations, labels
- Inference assigns truth value to each variable

(Pujara et al., ISWC13)
NELL experiments:

Target Set

**Task:** Compute truth values of a target set derived from the evaluation data

**Comparisons:**

- **Baseline:** Average confidences of extractors for each fact in the NELL candidates
- **NELL:** Evaluate NELL’s promotions (on the full knowledge graph)
- **MLN:** Method of (Jiang, ICDM12) – estimates marginal probabilities with MC-SAT
- **PSL-KGI:** Apply full Knowledge Graph Identification model

**Running Time:** Inference completes in 10 seconds, values for 25K facts

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.873</td>
<td>.828</td>
</tr>
<tr>
<td>NELL</td>
<td>.765</td>
<td>.673</td>
</tr>
<tr>
<td>MLN (Jiang, ICDM12)</td>
<td>.899</td>
<td>.836</td>
</tr>
<tr>
<td>PSL-KGI</td>
<td>.904</td>
<td>.853</td>
</tr>
</tbody>
</table>
NELL experiments:
Complete knowledge graph

Task: Compute a full knowledge graph from uncertain extractions

Comparisons:

NELL  NELL’s strategy: ensure ontological consistency with existing KB
PSL-KGI  Apply full Knowledge Graph Identification model

Running Time: Inference completes in 130 minutes, producing 4.3M facts

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NELL</td>
<td>0.765</td>
<td>0.801</td>
<td>0.477</td>
<td>0.634</td>
</tr>
<tr>
<td>PSL-KGI</td>
<td>0.892</td>
<td>0.826</td>
<td>0.871</td>
<td>0.848</td>
</tr>
</tbody>
</table>
Ontology-Aware Partitioning
Problem: Partition the Knowledge Graph
Key idea: Ontology-aware partitioning

- Partition the *ontology* graph, not the knowledge graph

- Induce a partitioning of the knowledge graph based on the ontology partition

(Pujara et al., AKBC13)
Experiments: Scalability

Partitions vs. Performance

- Running Time (minutes)
- Number of Partitions
- Area Under Precision-Recall Curve

Graph showing the relationship between the number of partitions and performance metrics such as running time and area under the precision-recall curve.
Dynamic Knowledge Graphs
Problem: Incremental Updates to KG

How do we add new extractions to the Knowledge Graph?
Naïve Approach: Full KGI over extractions
Approximation: KGI over subset of graph
Conclusion

- Knowledge Graph Identification is a powerful technique for producing knowledge graphs from noisy IE system output

- Using PSL we are able to enforce global ontological constraints and capture uncertainty in our model

- Unlike previous work, our approach infers complete knowledge graphs for datasets with millions of extractions

Code available on GitHub:

https://github.com/linqs/KnowledgeGraphIdentification

Questions?