

KNOWLEDGE GRAPH IDENTIFICATION

Jay Pujara¹, Hui Miao¹, Lise Getoor¹, William Cohen²

¹University of Maryland, College Park, US

²Carnegie Mellon University

International Semantic Web Conference
10/25/2013



Overview

Problem:

Build a Knowledge Graph from millions of noisy extractions

Approach:

Knowledge Graph Identification reasons jointly over all facts in the knowledge graph

Method:

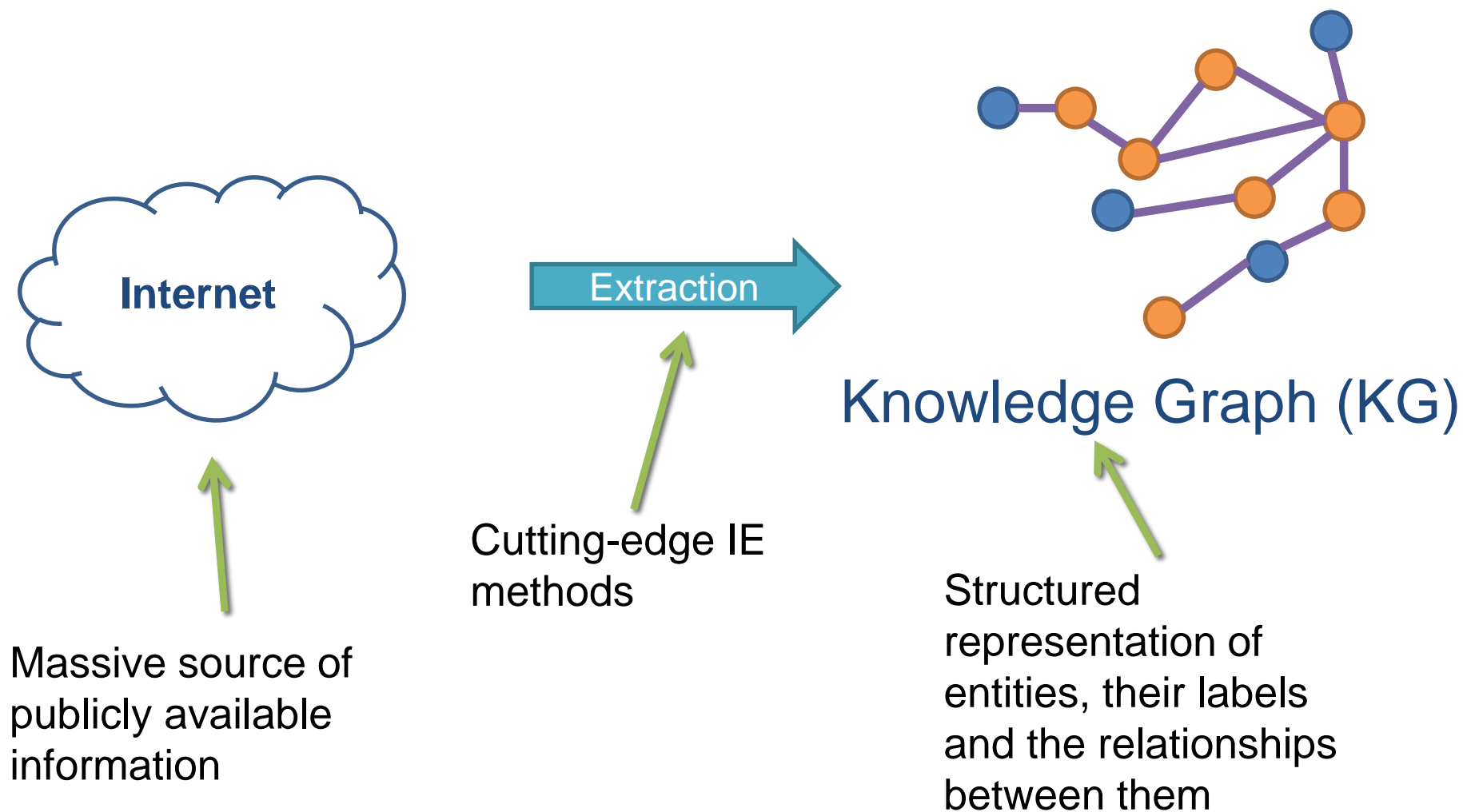
Use probabilistic soft logic to easily specify models and efficiently optimize them

Results:

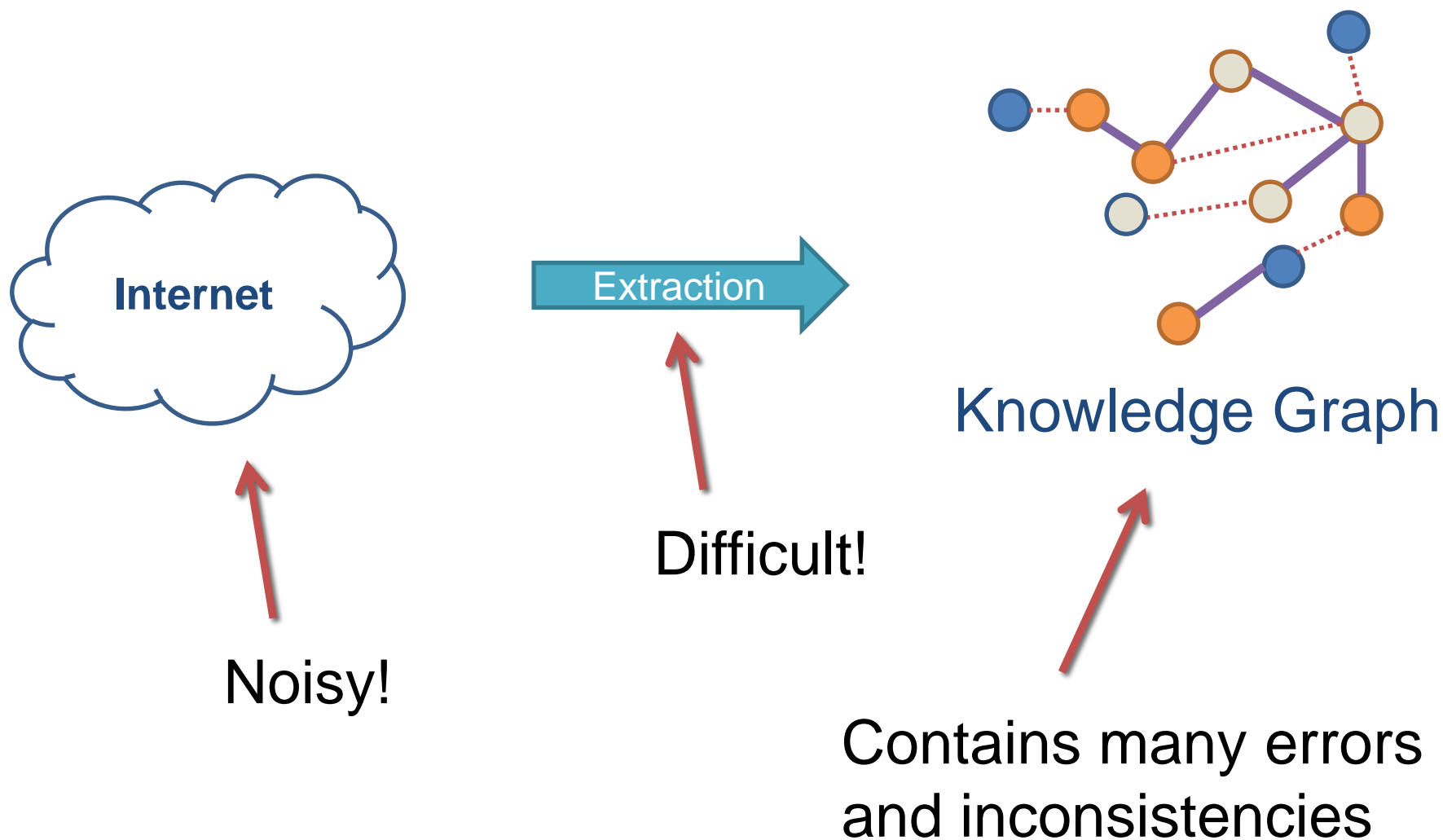
State-of-the-art performance on real-world datasets producing knowledge graphs with millions of facts

CHALLENGES IN KNOWLEDGE GRAPH CONSTRUCTION

Motivating Problem: New Opportunities



Motivating Problem: Real Challenges



NELL: The Never-Ending Language Learner

NELL @cmunell 1 Oct
True or False? "kevn tv" is a #TVStation (bit.ly/18JQ8gs)
Expand

NELL @cmunell 1 Oct
True or False? "metro-Atlanta" is a #County (bit.ly/1hhsefl)
Expand

NELL @cmunell 1 Oct
True or False? "exclusive right" is an #Artery (bit.ly/1bZq2LA)
Expand

NELL @cmunell 1 Oct
True or False? "Fireplace" is #SomethingFoundInOrOnBuildings (bit.ly/17E1JhW)
Expand Reply Retweet Favorite More

NELL @cmunell 1 Oct
True or False? "will_whalen" is an #AustralianPerson (bit.ly/1fUzRdT)
Expand

NELL @cmunell 30 Sep
True or False? "iron_chair" is a #HouseholdItem (bit.ly/14ZsCNk)
Expand

NELL @cmunell 30 Sep
True or False? "jerry gordon" is a #Chef (bit.ly/19Ry4QN)
Expand

- Large-scale IE project (Carlson et al., 2010)
- Lifelong learning: aims to “read the web”
- Ontology of known labels and relations
- Knowledge base contains millions of facts

- [person](#)
 - monarch
 - astronaut
 - personbylocation
 - personnorthamerica
 - personcanada
 - personus
 - politicianus
 - personmexico
 - personeurope
 - personaustralia
 - personafrica
 - personsouthamerica
 - personasia
 - personantarctica
- visualartist
- model
- scientist
- journalist
- female
- actor
- professor
- director
- architect
- politician
 - politicianus
- musician
- athlete
- chef
- male
- writer
- ceo
- judge
- mlauthor
- coach
- celebrity
- comedian
- criminal

Examples of NELL errors

Entity co-reference errors

Kyrgyzstan has many variants:

- Kyrgystan
- Kyrgistan
- Kyrghyzstan
- Kyrgyzstan
- 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.

[Home](#) > [Holiday Destinations](#) > [Kyrghyzstan](#) > [Bishkek](#) > [Climate Profile](#)



Fast Forecast

Holiday Weather

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

Missing and spurious labels

[Anssi Kullberg](#) has sent along some great trip reports to unusual places, including [Kyrgyzstan](#), [Pakistan](#), [Egypt/Jordan](#), and [Afghanistan](#). I had to create a whole new country page for [Afghanistan](#) to hold that last one! Thanks so much, Anssi!

[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](#).

Kyrgyzstan ([/kɜrɡɪˈstɑːn/](#) *kur-gi-STANN*;^[5] [Kyrgyz](#): Кыргызстан (IPA: [qɯrʁwɯsˈstɑn]); [Russian](#): Киргизия), officially the **Kyrgyz Republic** ([Kyrgyz](#): Кыргыз Республикасы; [Russian](#): Кыргызская Республика), is a **country** located in [Central Asia](#).^[6] 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 is
labeled a bird and
a country

Missing and spurious relations

Guidance

Kazakhstan / Kyrgyzstan – Consular Fees

Organisation: [Foreign & Commonwealth Office](#)
Page history: [Published 4 April 2013](#)

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

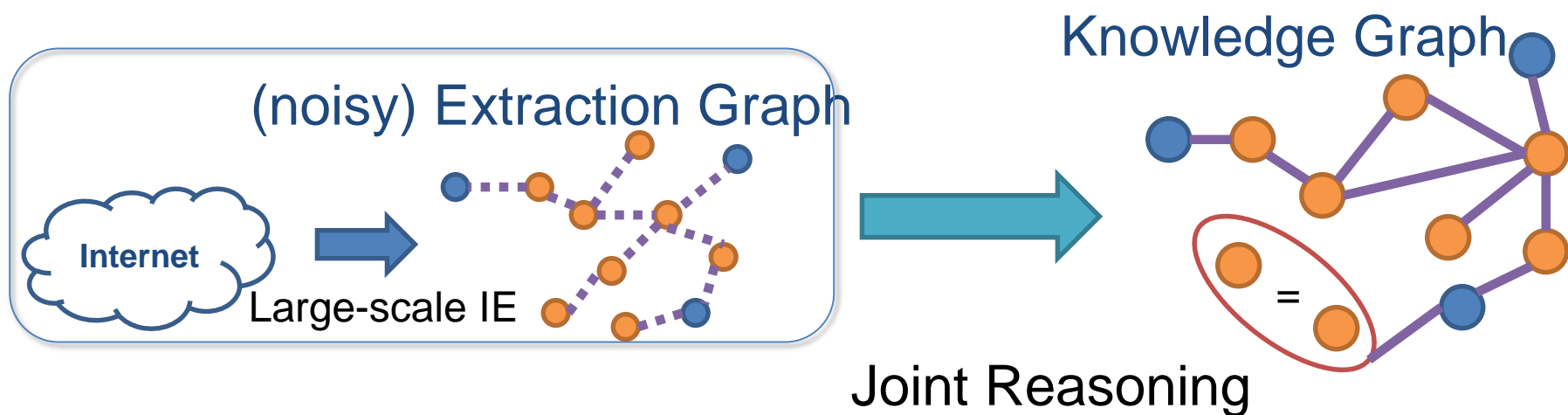
Violations of ontological knowledge

- 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 require **jointly** considering multiple extractions

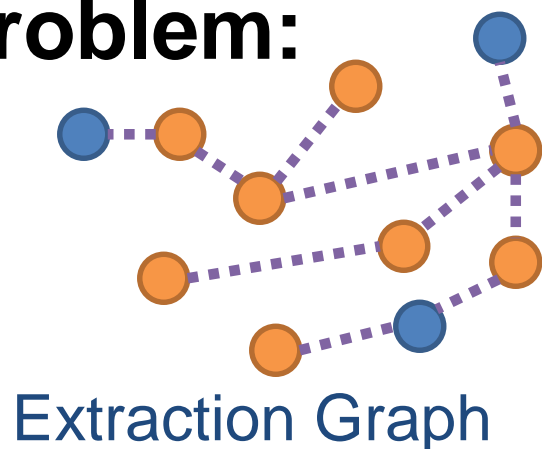
KNOWLEDGE GRAPH IDENTIFICATION

Motivating Problem (revised)



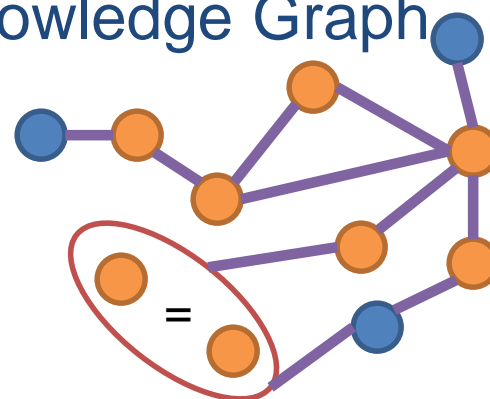
Knowledge Graph Identification

Problem:



Knowledge
Graph
Identification

Knowledge Graph



Solution: *Knowledge Graph Identification (KGI)*

- Performs *graph identification*:
 - entity resolution
 - collective classification
 - 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: Extraction Graph

Uncertain Extractions:

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

Extraction Graph

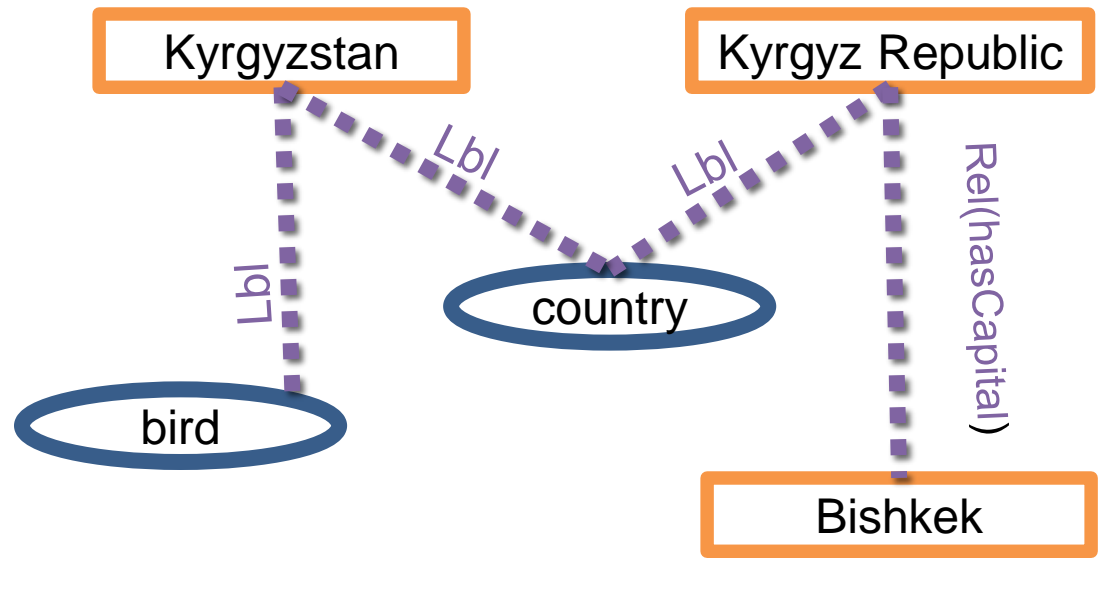


Illustration of KGI: Ontology + ER

Uncertain Extractions:

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

Ontology:

- Dom(hasCapital, country)
- Mut(country, bird)

Entity Resolution:

- SameEnt(Kyrgyz Republic, Kyrgyzstan)

(Annotated) Extraction Graph

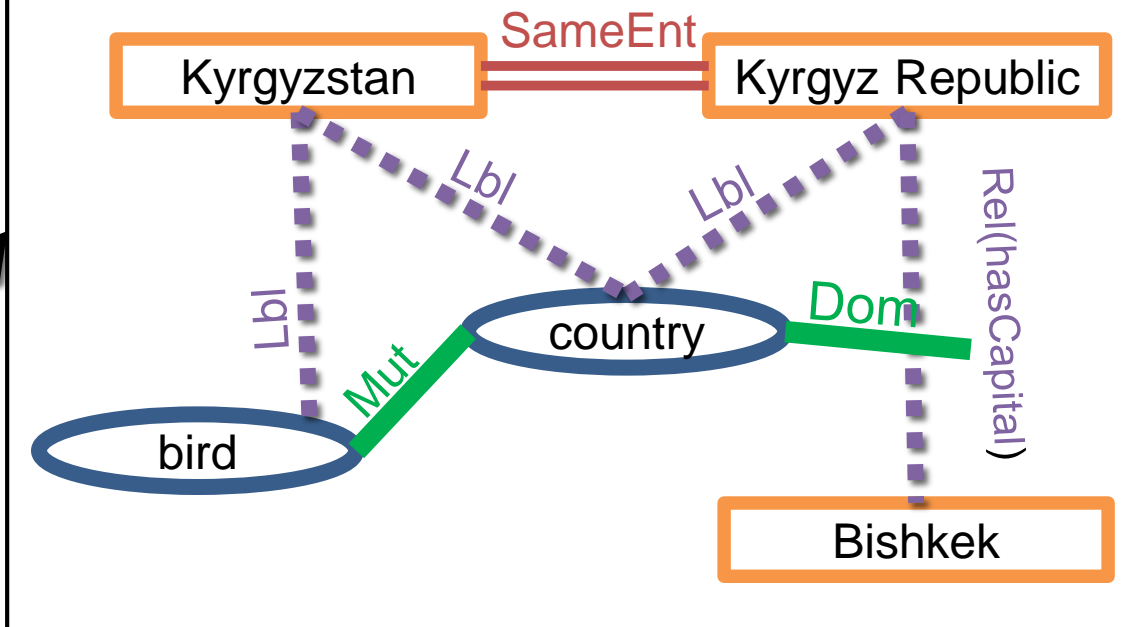


Illustration of KGI

Uncertain Extractions:

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

Ontology:

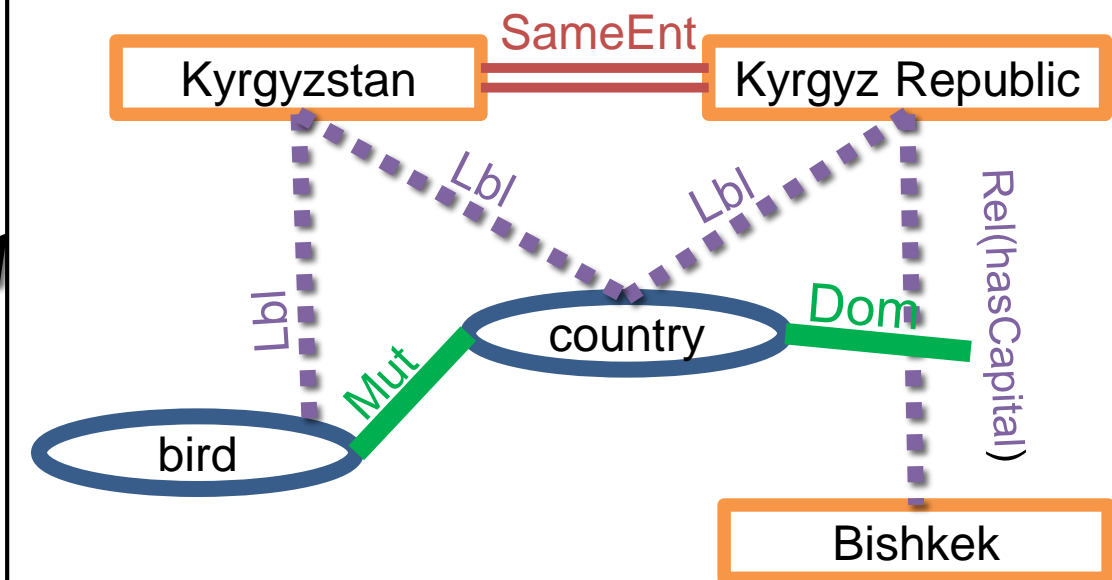
Dom(hasCapital, country)

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

Lbl(Kyrgyzstan,
bird)

Rel(hasCapital,
Kyrgyzstan,
Bishkek)

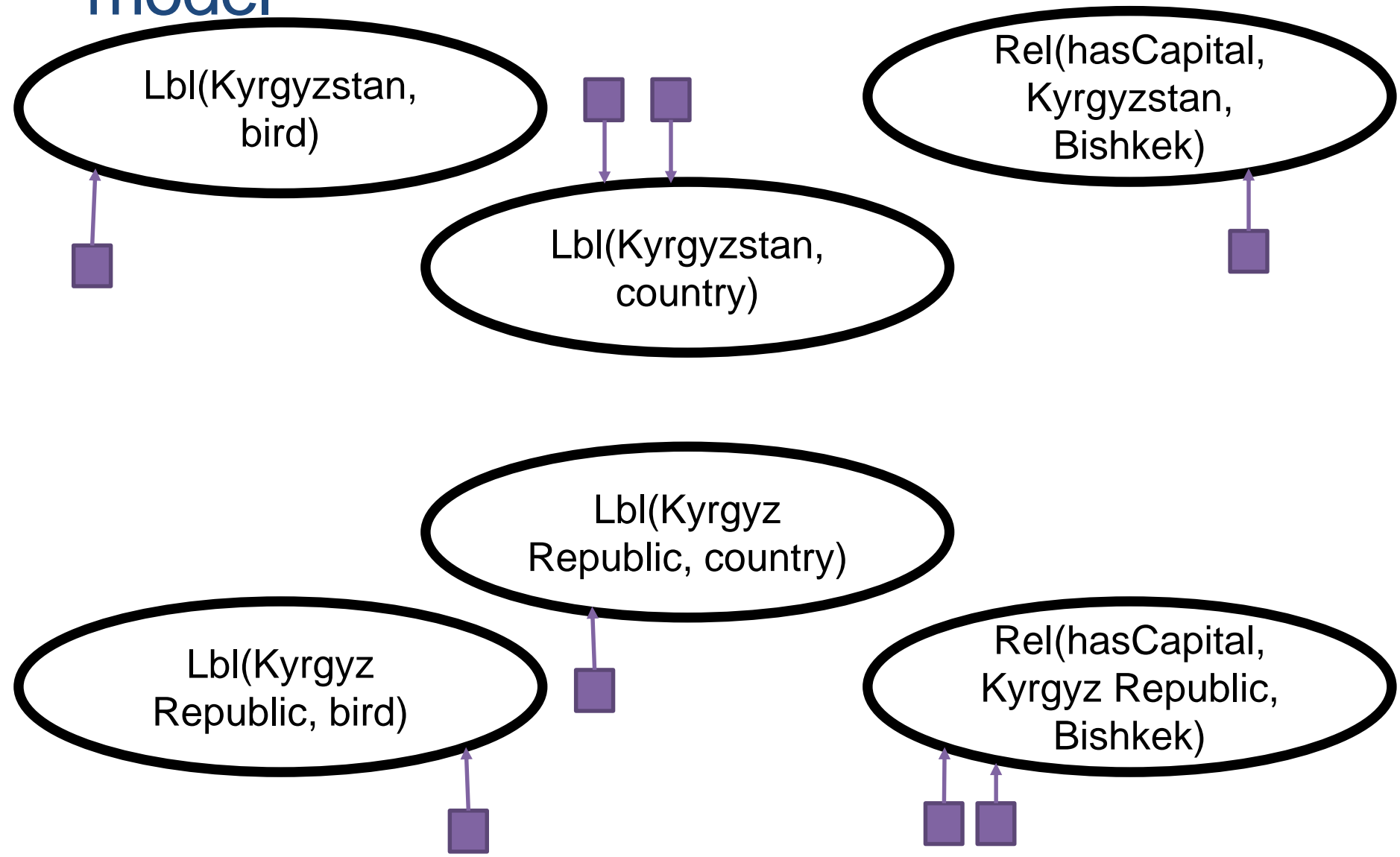
Lbl(Kyrgyzstan,
country)

Lbl(Kyrgyz
Republic, country)

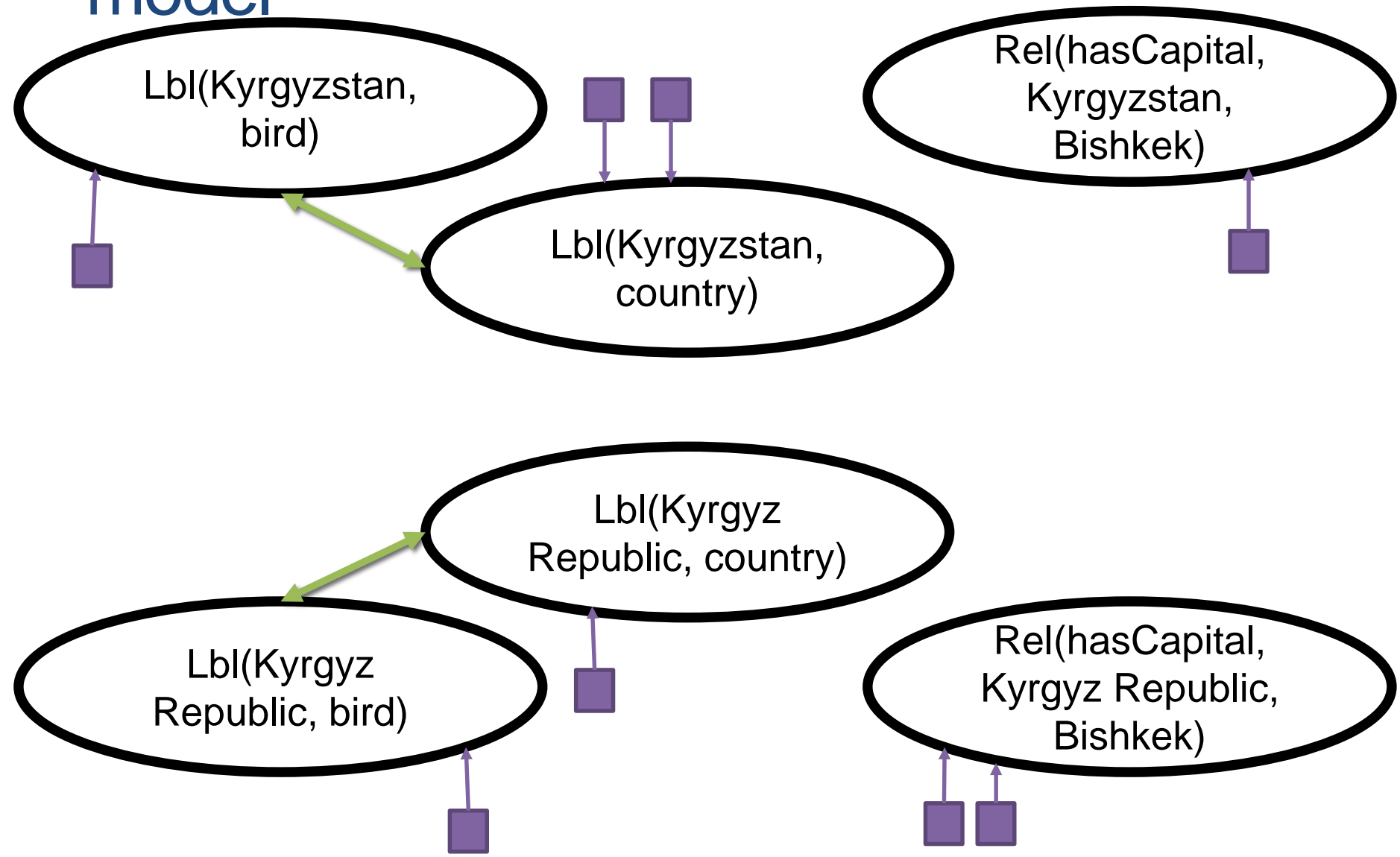
Lbl(Kyrgyz
Republic, bird)

Rel(hasCapital,
Kyrgyz Republic,
Bishkek)

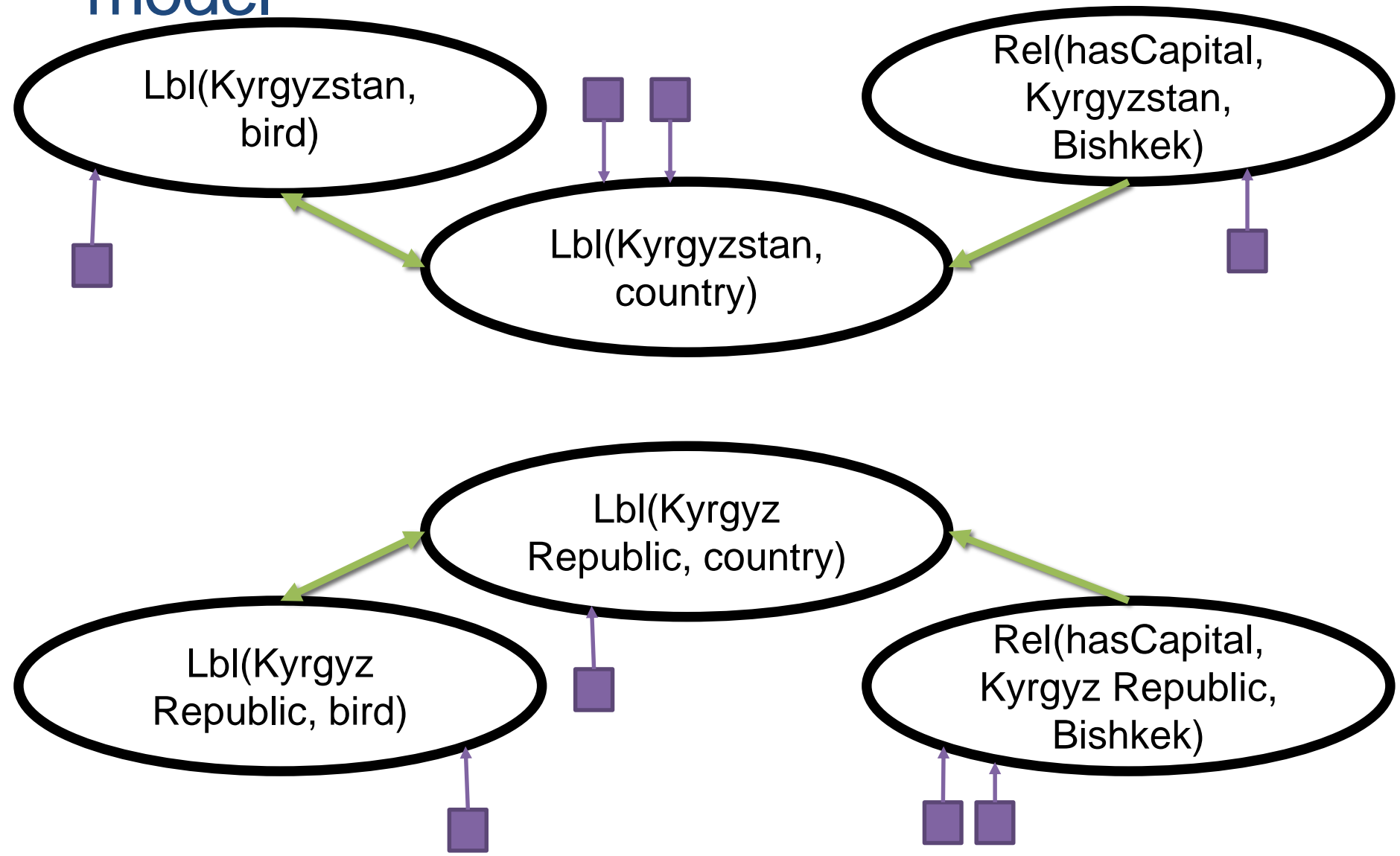
Viewing KGI as a probabilistic graphical model



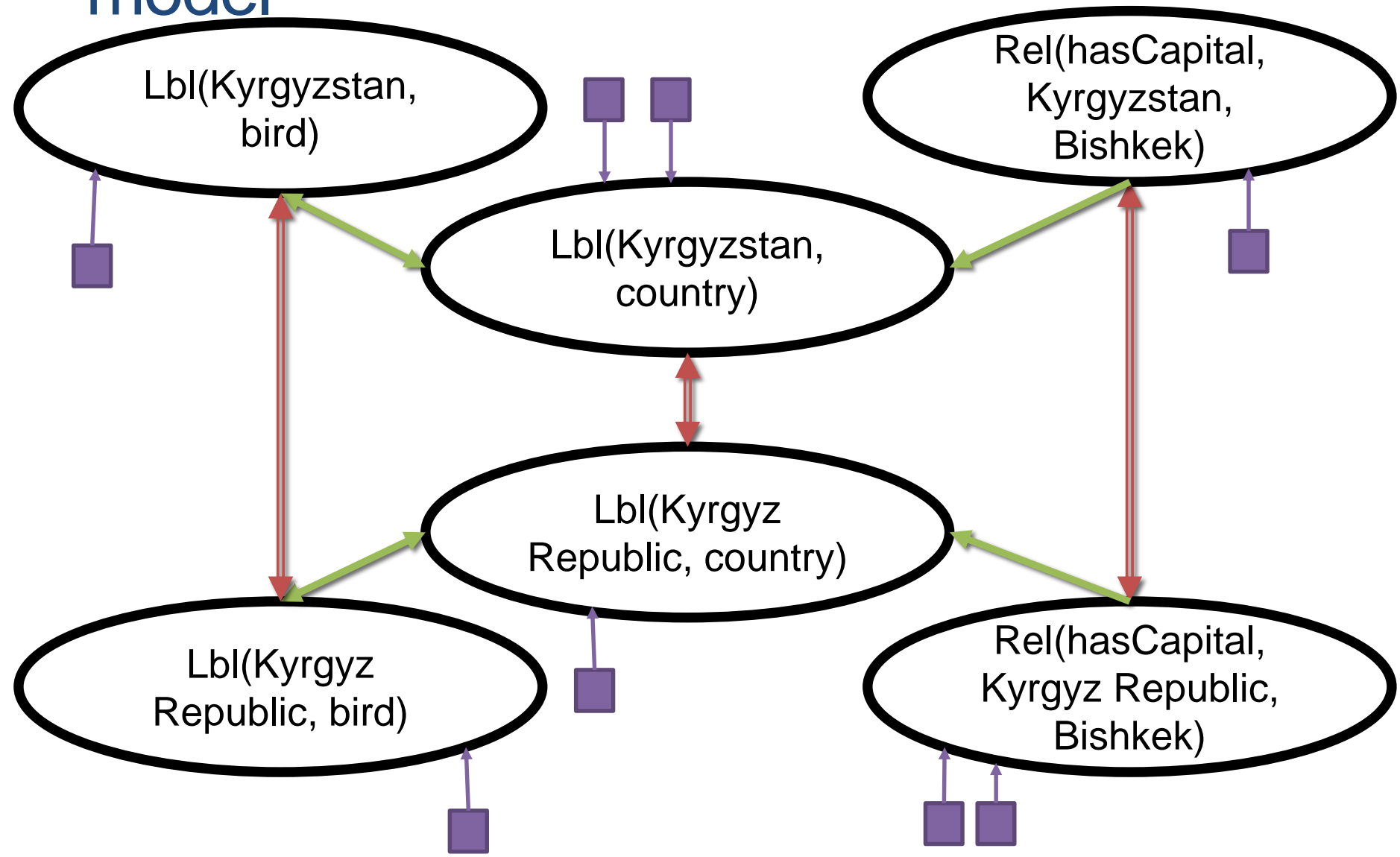
Viewing KGI as a probabilistic graphical model



Viewing KGI as a probabilistic graphical model

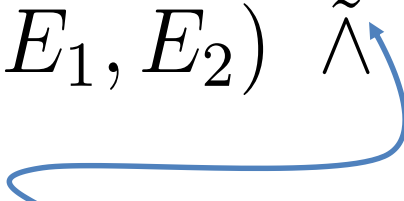


Viewing KGI as a probabilistic graphical model



Background: Probabilistic Soft Logic (PSL)

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

$$\text{SAMEENT}(E_1, E_2) \tilde{\wedge} \text{LBL}(E_1, L) \Rightarrow \text{LBL}(E_2, L)$$


- 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

$w_{EL} : \text{SAMEENT}(\text{Kyrgyzstan}, \text{Kyrgyz Republic}) \tilde{\wedge}$

$\text{LBL}(\text{Kyrgyzstan}, \text{country}) \Rightarrow \text{LBL}(\text{Kyrgyz Republic}, \text{country})$

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

$$P(G | E) = \frac{1}{Z} \exp\left(-\sum_{r \in R} w_r J_r(G)\right)$$

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

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 scales with $O(|R|)$

PSL Rules for the KGI Model

PSL Rules: Uncertain Extractions



$\mathbf{w_{CR-T}} : \mathbf{CANDREL}_T(E_1, E_2, R) \Rightarrow \mathbf{REL}(E_1, E_2, R)$

$\mathbf{w_{CL-T}} : \mathbf{CANDLBL}_T(E, L) \Rightarrow \mathbf{LBL}(E, L)$

PSL Rules: Uncertain Extractions

$\mathbf{w}_{\mathbf{CR-T}} : \mathbf{CANDREL}_T(E_1, E_2, R)$

$\mathbf{w}_{\mathbf{CL-T}} : \mathbf{CANDLBL}_T(E, L)$

Relation in
Knowledge Graph

 $\Rightarrow \mathbf{REL}(E_1, E_2, R)$
 $\Rightarrow \mathbf{LBL}(E, L)$

Label in
Knowledge Graph

PSL Rules: Uncertain Extractions

Predicate representing uncertain
relation extraction from extractor
T

$\mathbf{w_{CR-T}} : \mathbf{CANDREL}_T(E_1, E_2, R)$

$\mathbf{w_{CL-T}} : \mathbf{CANDLBL}_T(E, L)$

Predicate representing uncertain
label extraction from extractor T

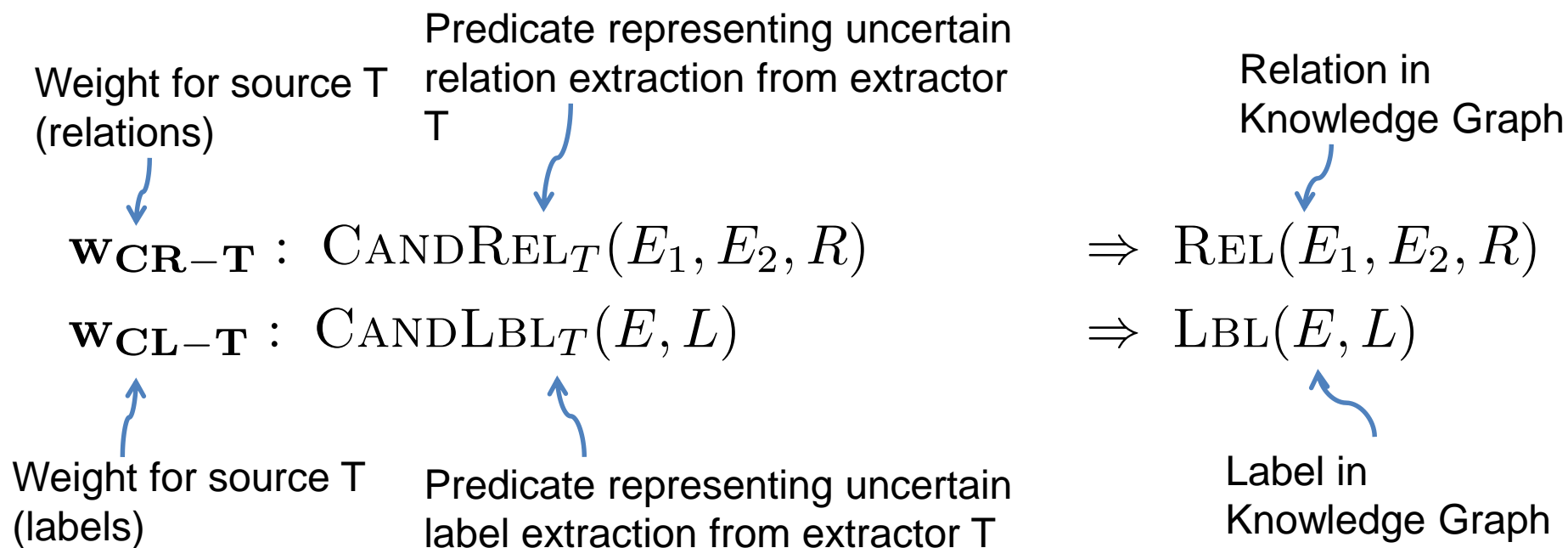
Relation in
Knowledge Graph

$\Rightarrow \mathbf{REL}(E_1, E_2, R)$

$\Rightarrow \mathbf{LBL}(E, L)$

Label in
Knowledge Graph

PSL Rules: Uncertain Extractions



PSL Rules: Entity Resolution

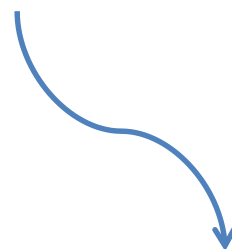
$$w_{EL} : \text{SAMEENT}(E_1, E_2) \tilde{\wedge} \text{LBL}(E_1, L) \Rightarrow \text{LBL}(E_2, L)$$

$$w_{ER} : \text{SAMEENT}(E_1, E_2) \tilde{\wedge} \text{REL}(E_1, E, R) \Rightarrow \text{REL}(E_2, E, R)$$

$$w_{ER} : \text{SAMEENT}(E_1, E_2) \tilde{\wedge} \text{REL}(E, E_1, R) \Rightarrow \text{REL}(E, E_2, R)$$



ER 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

PSL Rules: Ontology

Inverse:

$$\mathbf{w}_O : \text{INV}(R, S) \quad \tilde{\wedge} \text{REL}(E_1, E_2, R) \Rightarrow \text{REL}(E_2, E_1, S)$$

Selectional Preference:

$$\mathbf{w}_O : \text{DOM}(R, L) \quad \tilde{\wedge} \text{REL}(E_1, E_2, R) \Rightarrow \text{LBL}(E_1, L)$$

$$\mathbf{w}_O : \text{RNG}(R, L) \quad \tilde{\wedge} \text{REL}(E_1, E_2, R) \Rightarrow \text{LBL}(E_2, L)$$

Subsumption:

$$\mathbf{w}_O : \text{SUB}(L, P) \quad \tilde{\wedge} \text{LBL}(E, L) \Rightarrow \text{LBL}(E, P)$$

$$\mathbf{w}_O : \text{RSUB}(R, S) \quad \tilde{\wedge} \text{REL}(E_1, E_2, R) \Rightarrow \text{REL}(E_1, E_2, S)$$

Mutual Exclusion:

$$\mathbf{w}_O : \text{MUT}(L_1, L_2) \quad \tilde{\wedge} \text{LBL}(E, L_1) \Rightarrow \tilde{\neg} \text{LBL}(E, L_2)$$

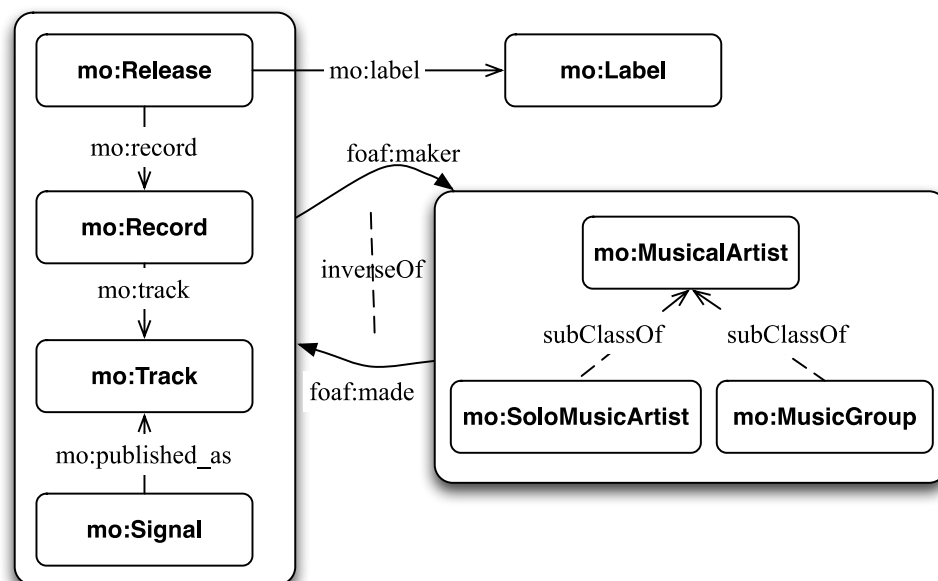
$$\mathbf{w}_O : \text{RMUT}(R, S) \quad \tilde{\wedge} \text{REL}(E_1, E_2, R) \Rightarrow \tilde{\neg} \text{REL}(E_1, E_2, S)$$

EVALUATION

Two Evaluation Datasets

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

LinkedBrainz dataset for KGI



Mapping to FRBR/FOAF ontology

DOM	rdfs:domain
RNG	rdfs:range
INV	owl:inverseOf
SUB	rdfs:subClassOf
RSUB	rdfs:subPropertyOf
MUT	owl:disjointWith

Adding noise to LinkedBrainz

Add realistic noise to LinkedBrainz data:

Error Type	Erroneous Data
Co-reference	User misspells artist
Label	User swaps artist and album fields
Relation	User omits or adds spurious albums for artist
Reliability	Gaussian noise on truth value of information

LinkedBrainz experiments

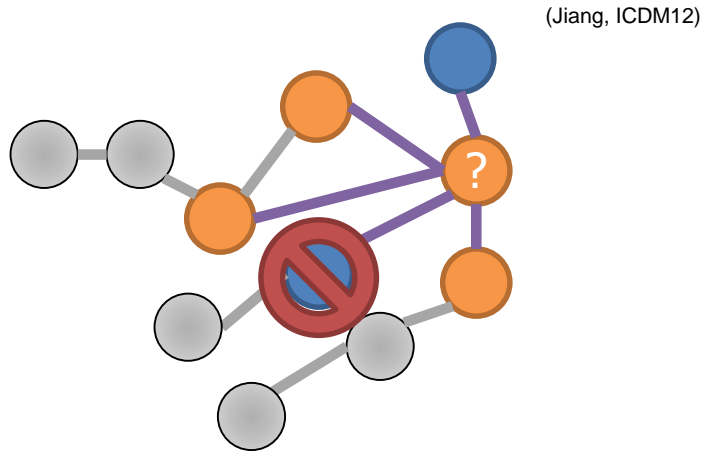
Comparisons:

- Baseline** Use noisy truth values as fact scores
- PSL-EROnly** Only apply rules for **Entity Resolution**
- PSL-OntOnly** Only apply rules for **Ontological reasoning**
- PSL-KGI model** Apply **Knowledge Graph Identification**

	AUC	Precision	Recall	F1 at .5	Max F1
Baseline	0.672	0.946	0.477	0.634	0.788
PSL-EROnly	0.797	0.953	0.558	0.703	0.831
PSL-OntOnly	0.753	0.964	0.605	0.743	0.832
PSL-KGI	0.901	0.970	0.714	0.823	0.919

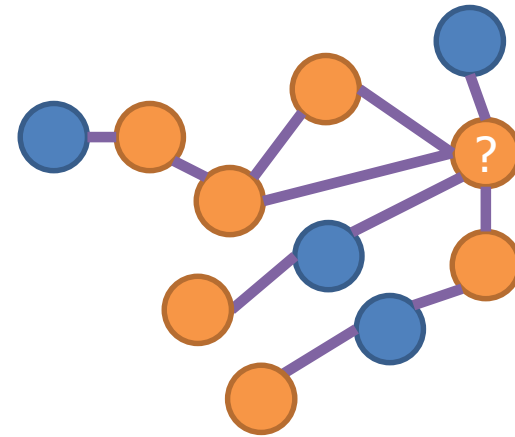
NELL Evaluation: two settings

Target Set: restrict to a subset of KG



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

Complete: Infer full knowledge graph



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

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

	AUC	F1
Baseline	.873	.828
NELL	.765	.673
MLN (Jiang, 12)	.899	.836
PSL-KGI	.904	.853

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

	AUC	Precision	Recall	F1
NELL	0.765	0.801	0.477	0.634
PSL-KGI	0.892	0.826	0.871	0.848

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?