UNIVERSITY OF TEXAS ARLINGTON

iDiR

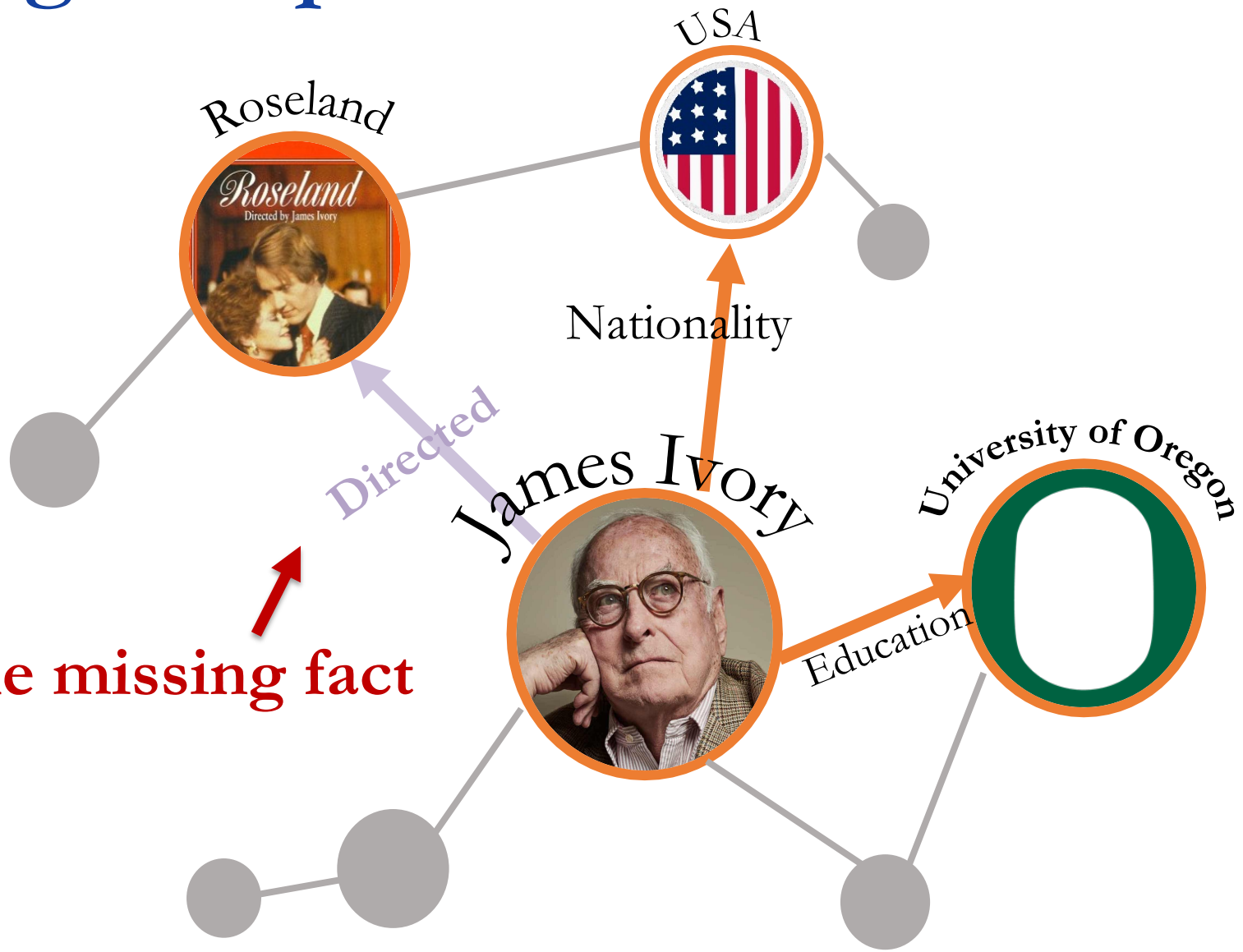
Comprehensive Analysis of Freebase and Dataset Creation for Robust Evaluation of Knowledge Graph Link Prediction Models

Nasim Shirvani-Mahdavi, Farahnaz Akrami, Mohammed Samiul Saeef, Xiao Shi, and Chengkai Li

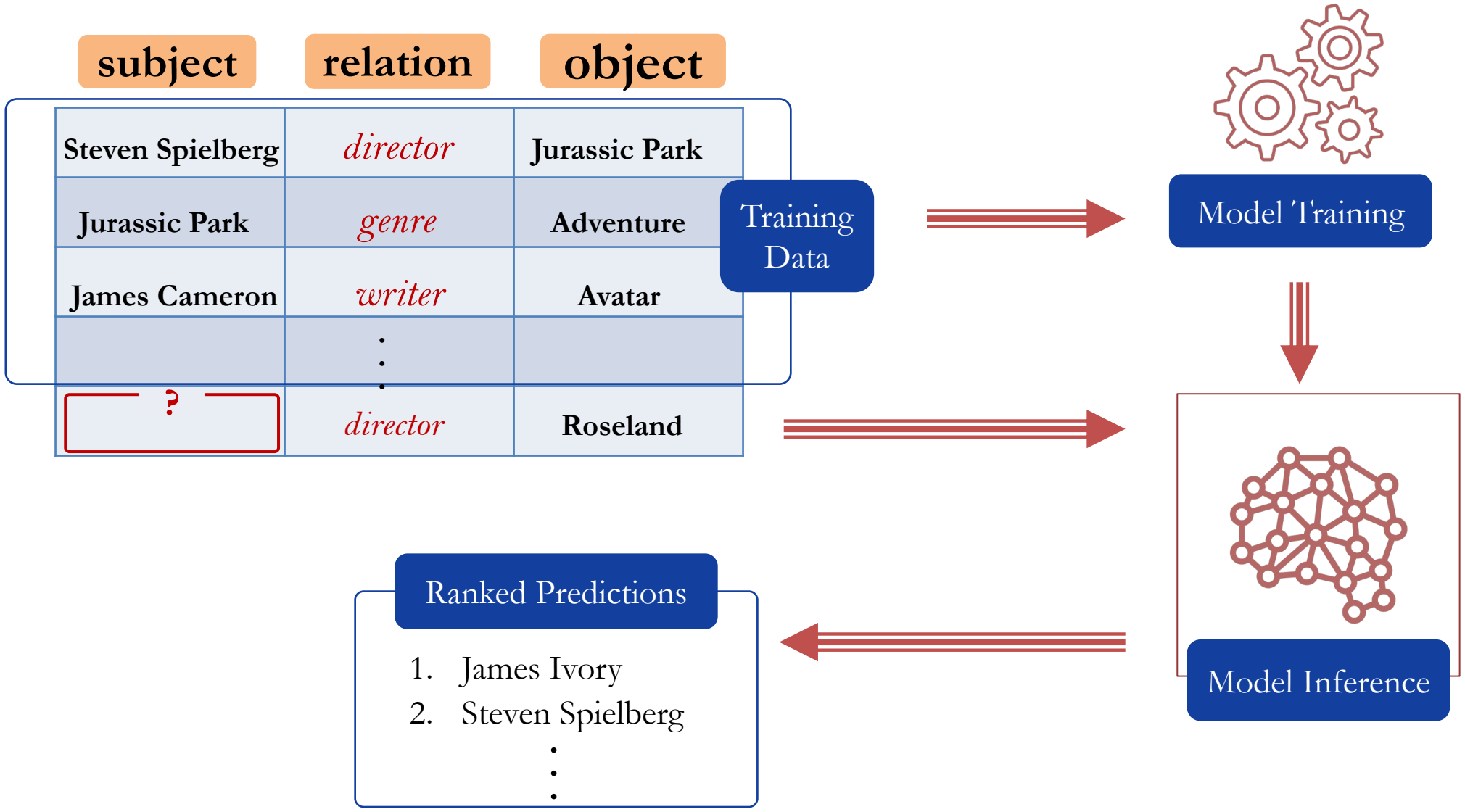
ISWC 2023 Resource Track

ISWC 2023
NOVEMBER 6-10
ATHENS-GREECE

Lack of Completeness Hinders Applications of Knowledge Graphs



Machine Learning Models for Link Prediction



Evaluation of Embedding Models

- **Evaluation metrics**
  - MR, FMR ⬇
  - MRR, FMRR ⬆
  - Hits@k, FHits@k ⬆
- **Benchmark datasets**

  Freebase, Wikidata, WordNet, YAGO, NELL, DBpedia, …

Model	Year	Citation
TransE	2013	7373
DistMult	2015	2887
TransR	2015	1777
ComplEx	2016	2675
RotatE	2019	1722
…	…	…

How Do These Models Perform on Large Scale Datasets?

	340M triples	16M triples	500M triples
Model	Freebase86m	ogbl-wikikg2	WikiKG90M-LSC
TransE	.72	.52	.88
DistMult	.83	.37	.86
RotatE	.82	.49	.88

Includes non-subject matter triples

No multiary relationships

No multiary relationships

All measures are in FMRR

Large Realistic Datasets are Missing from Link Prediction Studies

Year 2022, 53 CSRankings publications on knowledge graph completion

- 48 publications used datasets from Freebase; only 3 used large-scale ones (Freebase86m)
- 8 publications used datasets from Wikidata; 5 used large-scale versions
- **All benchmark datasets either are problematic or small, or do not capture real-world data modeling idiosyncrasies**

12 top conferences in 2022: AAAL, IJCAI, WWW, KDD, ICML, ACL, EMNLP, NAACL, SIGIR, NeurIPS, SIGMOD, and VLDB

Model Performance Differ Drastically by Varying Dataset Size

Model	FB15K (DGL-KE)	FB15K-237 (DGL-KE)	FB15K-237 (LibKGE)	FB-CVT-REV (Large, DGL-KE)
TransE	.63	.24	.31	.67
DistMult	.68	.24	.34	.70
TransR	.66	.57	-	.66
ComplEx	.74	.23	.34	.71
RotatE	.68	.24	.33	.80

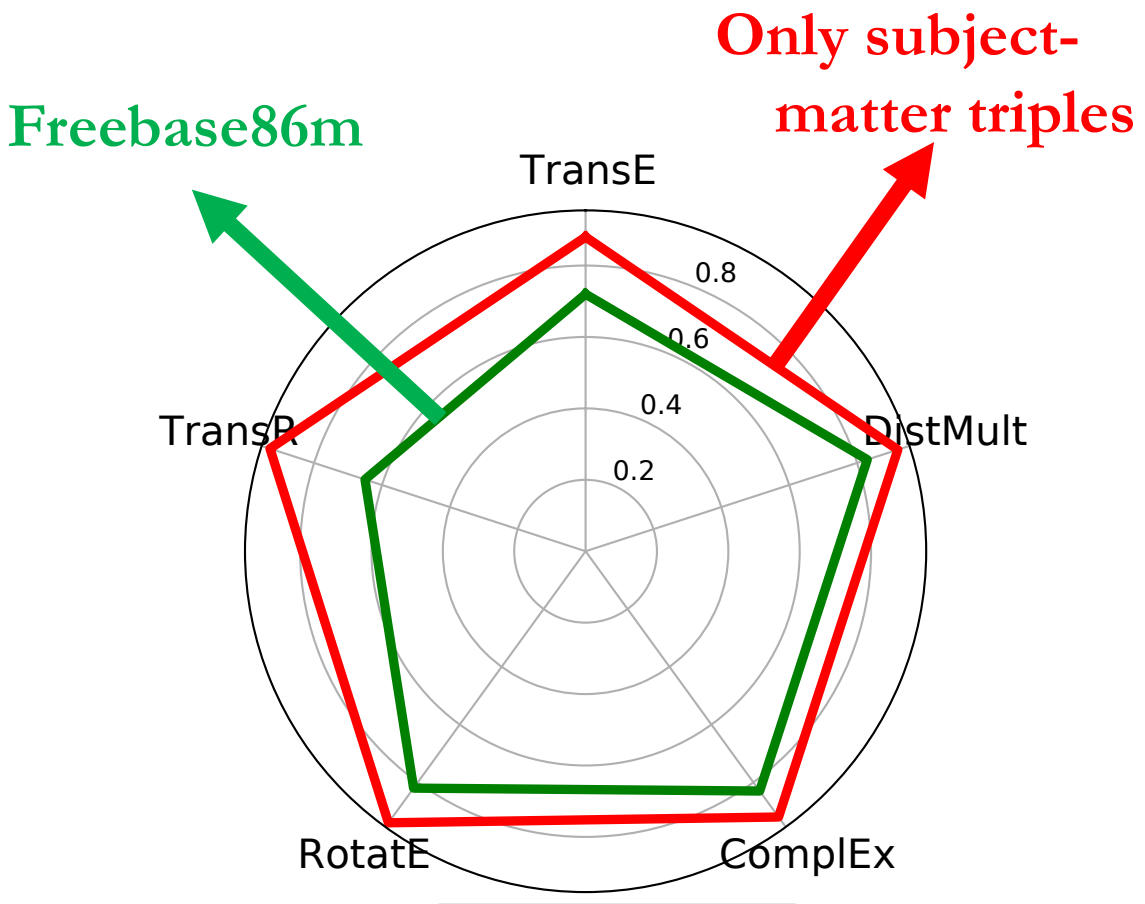
All measures are in FMRR

Contributions of Our Paper

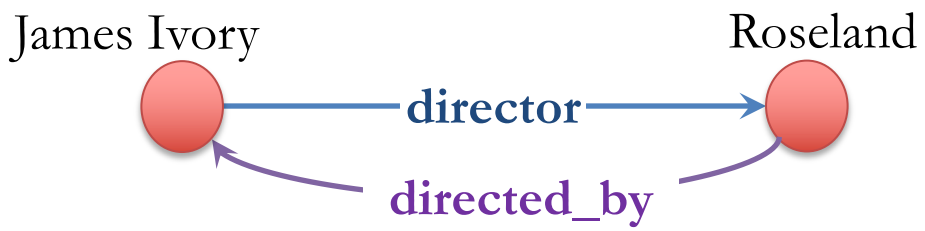
- Reported link prediction models' **true** performance on full-size Freebase datasets
- Reported how their performance is affected by various **data modeling idiosyncrasies**
  - Reverse triples
  - Multiary relationships Mediator nodes
  - Type system
- Made available **thoroughly prepared full-size Freebase datasets**

Mixing Non-Subject Matter Triples with Other Triples Degenerates Performance

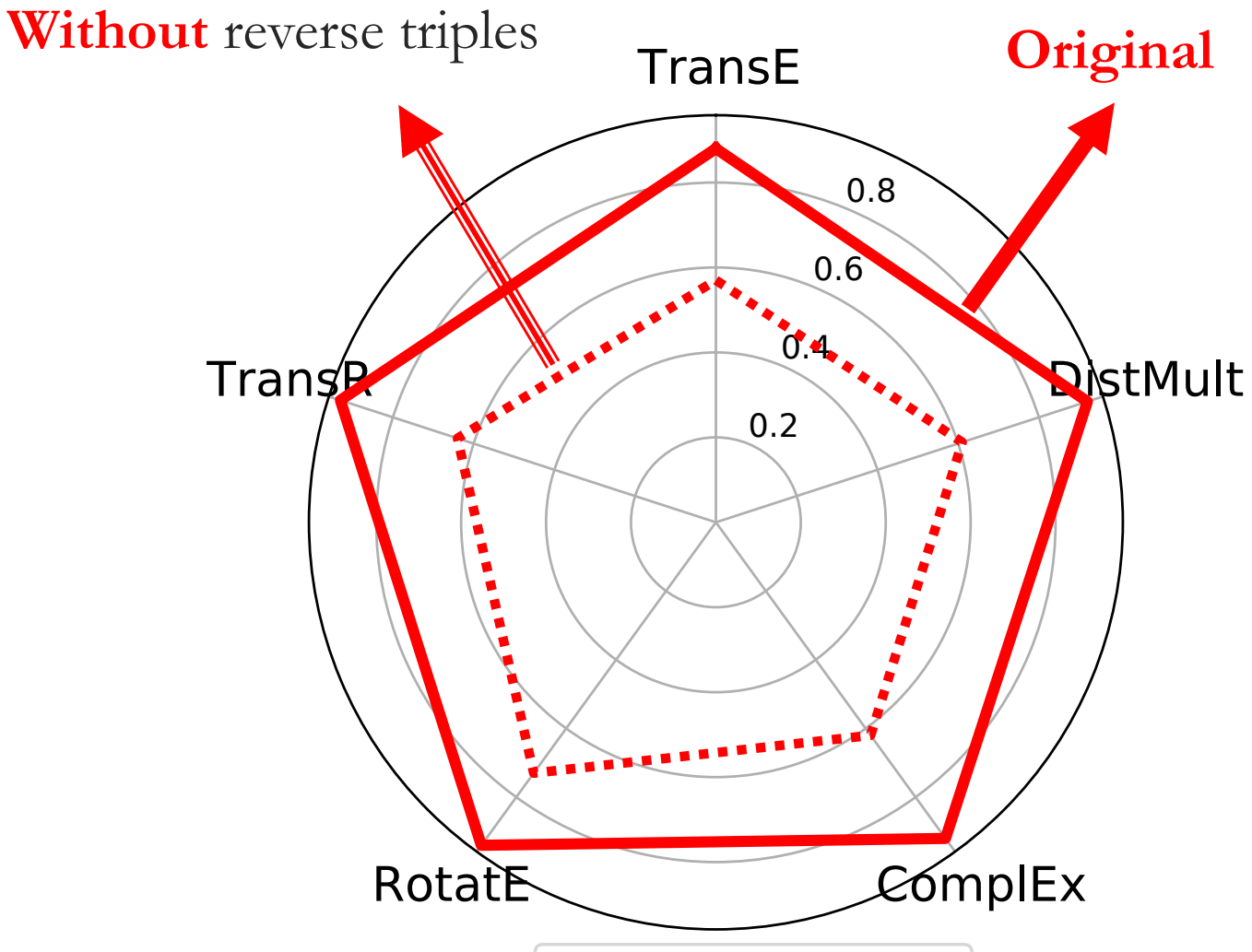
31% of the Freebase86m triples fall under non-subject matter domain. E.g., */dataworld/* and */freebase/*, two implementation domains



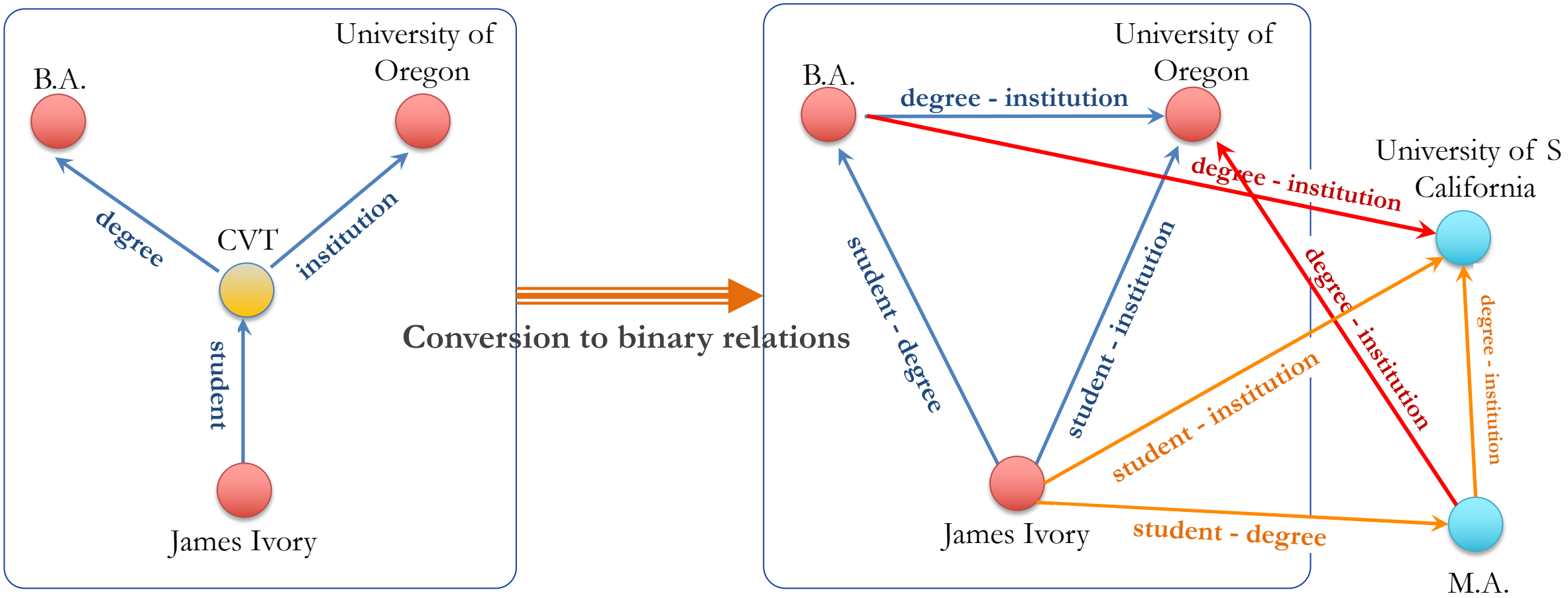
Reverse Triples in Freebase



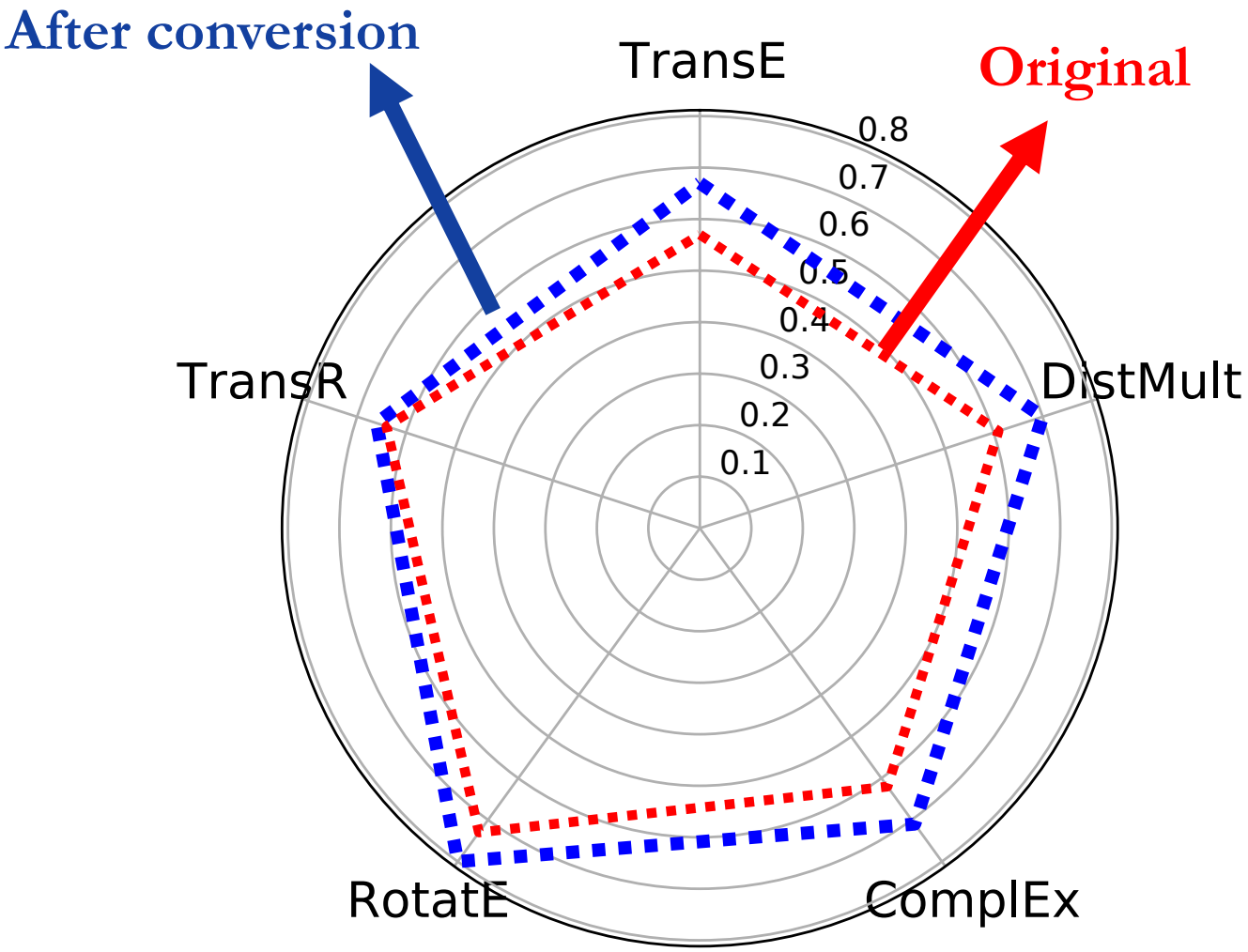
Reverse Triples Make LP Unrealistically Trivial



Complex Multiary Relationships Are Often Simplified to Binary Relations



Converting Multiary Relation to Binary Makes LP Easier

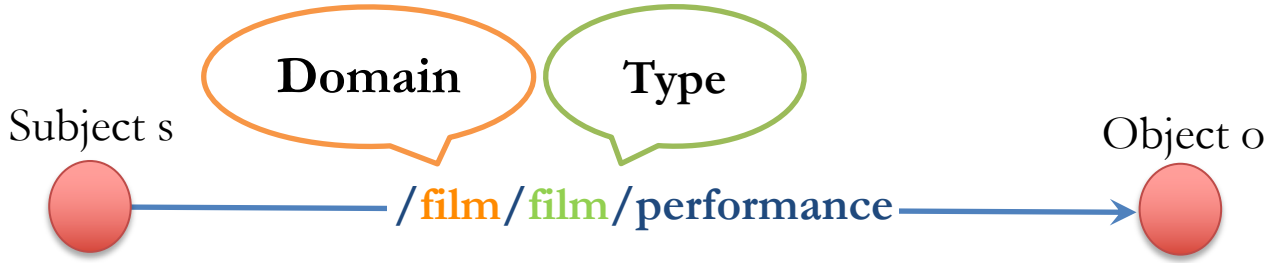


Converting Multiary Relation to Binary Makes LP Easier

	FB-CVT-REV			FB+CVT-REV		
Model	binary	concatenated	all	binary	multiary	all
TransE	.60	.90	.67	.57	.96	.57
DistMult	.64	.89	.70	.61	.77	.61
TransR	.58	.92	.66	.63	.87	.64
ComplEx	.66	.90	.71	.62	.80	.62
RotatE	.76	.92	.80	.73	.88	.73

All measures are in FMRR

Freebase Type System

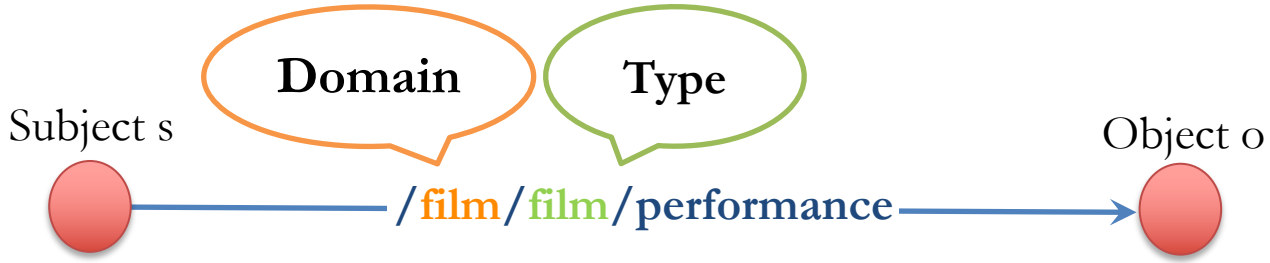


Object o belongs to types

P (o ∈ /film/actor) = 0.99
P (o ∈ /tv/tv_actor) = 0.10
P (o ∈ /music/artist) = 0.04
P (o ∈ /award/award_winner) = 0.03
P (o ∈ /people/person) = 0.99

*P is the probability of the object end of /film/film/performance belonging to type t

Freebase Type System



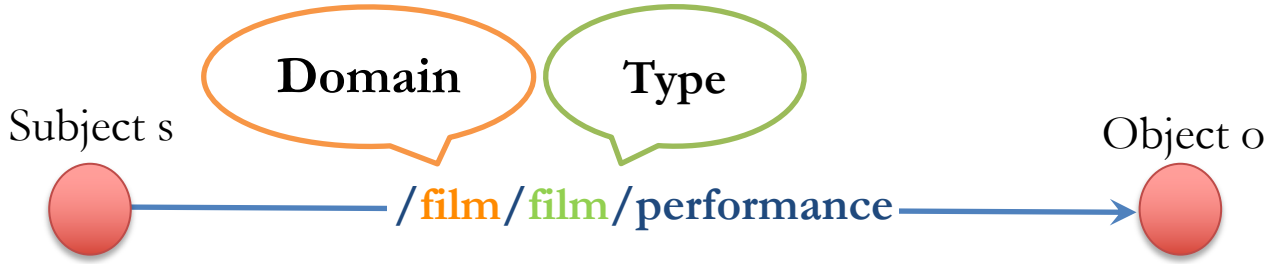
Object o belongs to types

- ✓ **P (o ∈/film/actor) = 0.99**
- ~~P (o ∈/tv/tv_actor) = 0.10~~
- ~~P (o ∈/music/artist) = 0.04~~
- ~~P (o ∈/award/award_winner) = 0.03~~
- ✓ **P (o ∈/people/person) = 0.99**

α = 0.95

*P is the probability of the object end of /film/film/performance belonging to type t

Freebase Type System



/people/person

P (o ∈ /film/actor | o ∈ /people/person) = 0.13

✓ **/film/actor**

P (o ∈ /people/person | o ∈ /film/actor) = **0.99**

Most specific type

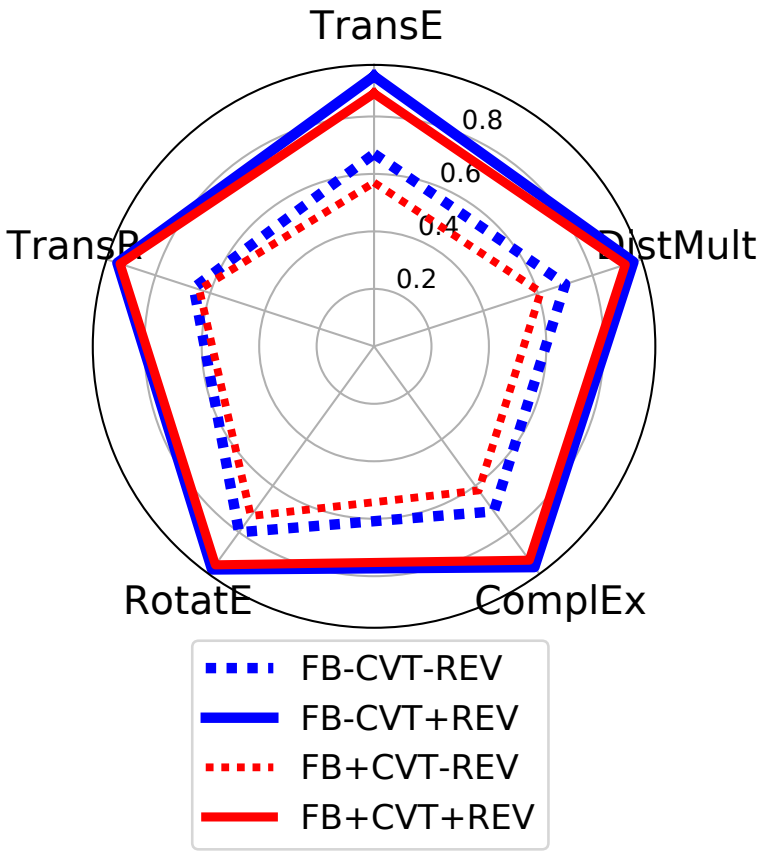
Usefulness of Freebase Type System in Creating Negative Samples

Task: Triple classification

Creating negative samples: Random triple corruption vs. Using type system

	consistent h		inconsistent h		consistent t		inconsistent t	
Model	accuracy	F1 score	accuracy	F1 score	accuracy	F1 score	accuracy	F1 score
TransE	.52	.55	.76	.74	.57	.56	.86	.86
DistMult	.53	.52	.91	.90	.58	.57	.92	.92
RotatE	.52	.52	.87	.86	.58	.53	.83	.82

Datasets Produced in This Work



Dataset	CVT	Reverse	#Entities	#Rels	#Triples
FB-CVT-REV	No	No	46M	3K	125M
FB-CVT+REV	No	Yes	46M	5K	238M
FB+CVT-REV	Yes	No	59M	2.6K	134M
FB+CVT+REV	Yes	Yes	59M	4.4K	244M

All measures are in FMRR

Take-Home: Contributions of Our Paper

- Reported link prediction models' **true** performance on full-size Freebase datasets

- Reported how their performance is affected by various **data modeling idiosyncrasies**
  - reverse triples
  - multiary relationships
  - type system

- Made available **thoroughly prepared full-size Freebase datasets**

Our Earlier Work and Resources

❖ This is a follow-up of our SIGMOD 2020 paper "Realistic re-evaluation of knowledge graph completion methods: An experimental study"

❖ GitHub repository of all source codes, datasets, and results

https://github.com/idirlab/freebases