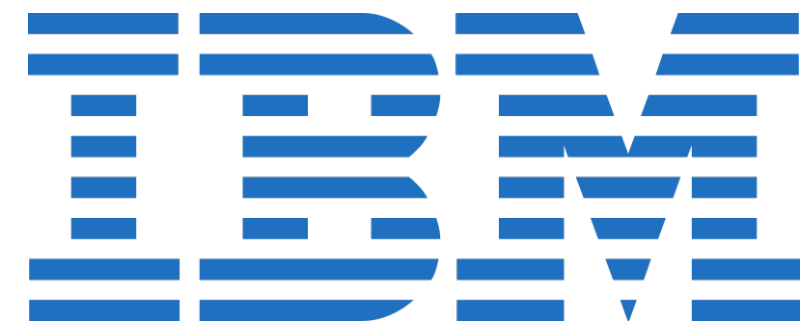


EmbedS: Scalable and Semantic-Aware Knowledge Graph Embeddings

Gonzalo I. Diaz¹, Achille Fokoue², Mohammad Sadoghi³



DEPARTMENT OF
**COMPUTER
SCIENCE**



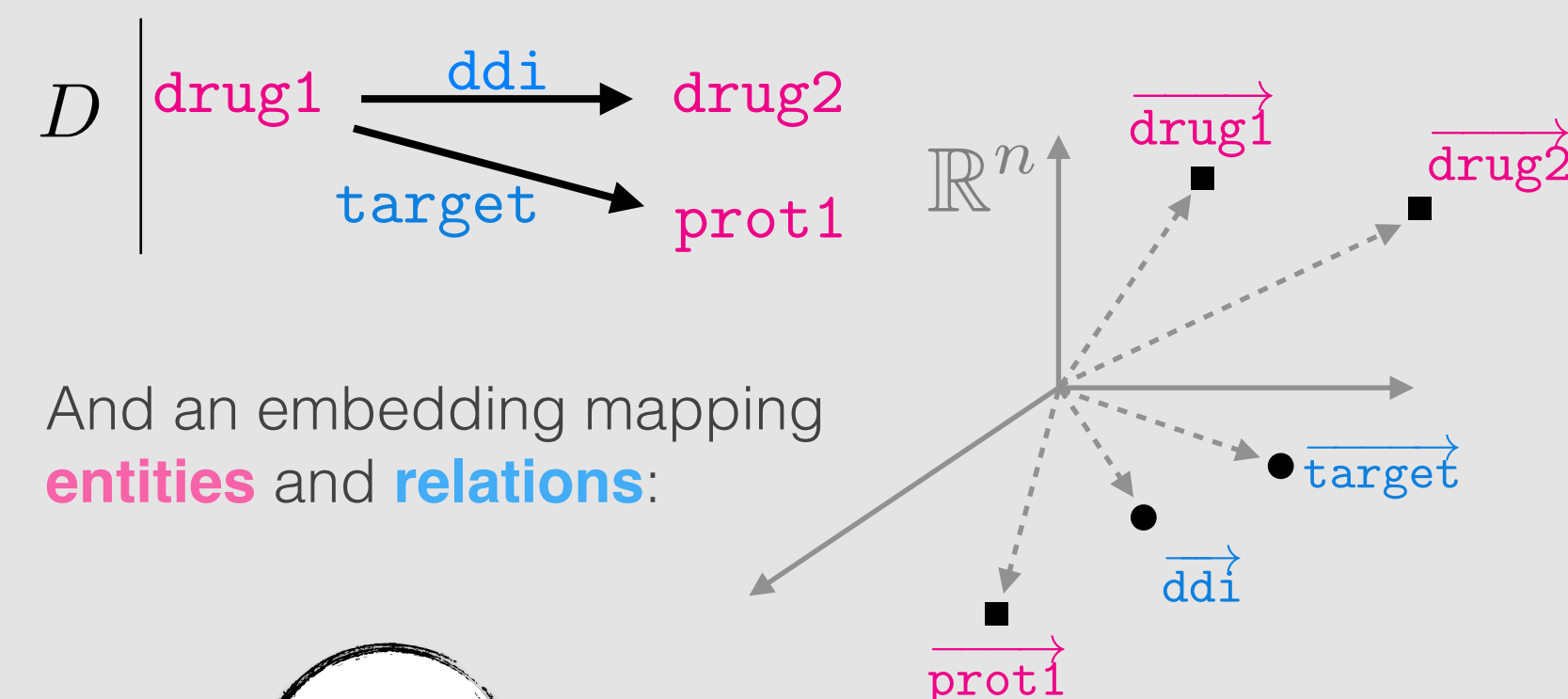
¹ University of Oxford
² IBM T.J. Watson Research Center
³ Exploratory Systems Lab
³ University of California, Davis



UC DAVIS
UNIVERSITY OF CALIFORNIA

Knowledge Graphs and Translational Embeddings

Assume a knowledge graph:



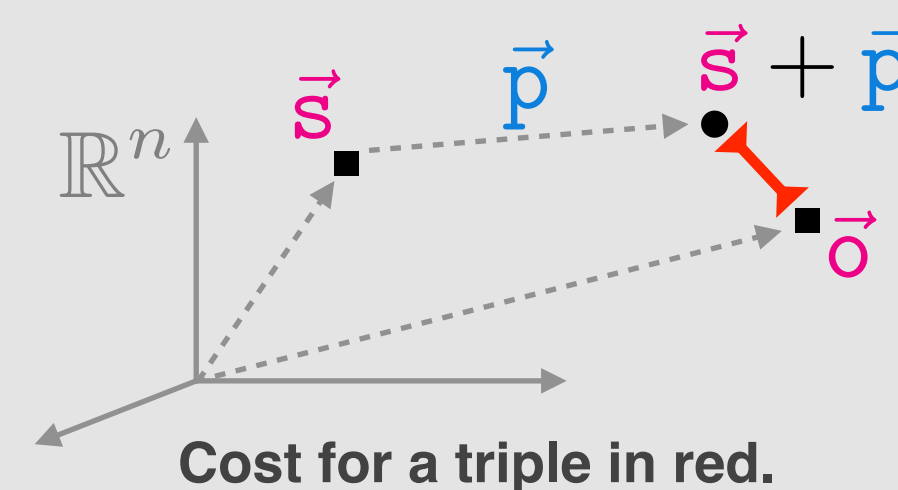
And an embedding mapping
entities and **relations**:

These models represent **relations** as a translational vector. Learning is achieved by minimizing the cost:

$$\mathcal{L} = \sum_{(s,p,o)} \left[\left| \left| (\vec{s} + \vec{p}) - \vec{o} \right| - \left| (\vec{s}' + \vec{p}) - \vec{o}' \right| + \gamma \right]_+$$

This model, **TransE** [1], was inspired by word2vec. The cost per triple is:

$$\mathcal{L}_{(s,p,o)} = \left| \left| (\vec{s} + \vec{p}) - \vec{o} \right| \right|$$



1

Ontology-aware embeddings

Ontology-unaware embeddings may violate semantics:

(drug1, target, prot1)	(uri1, uri2, uri3)
(prot1, rdf:type, Prot)	(uri3, uri4, uri5)
(target, rdfs:range, Prot)	(uri2, uri6, uri5)
Ontology-aware	Ontology-unaware

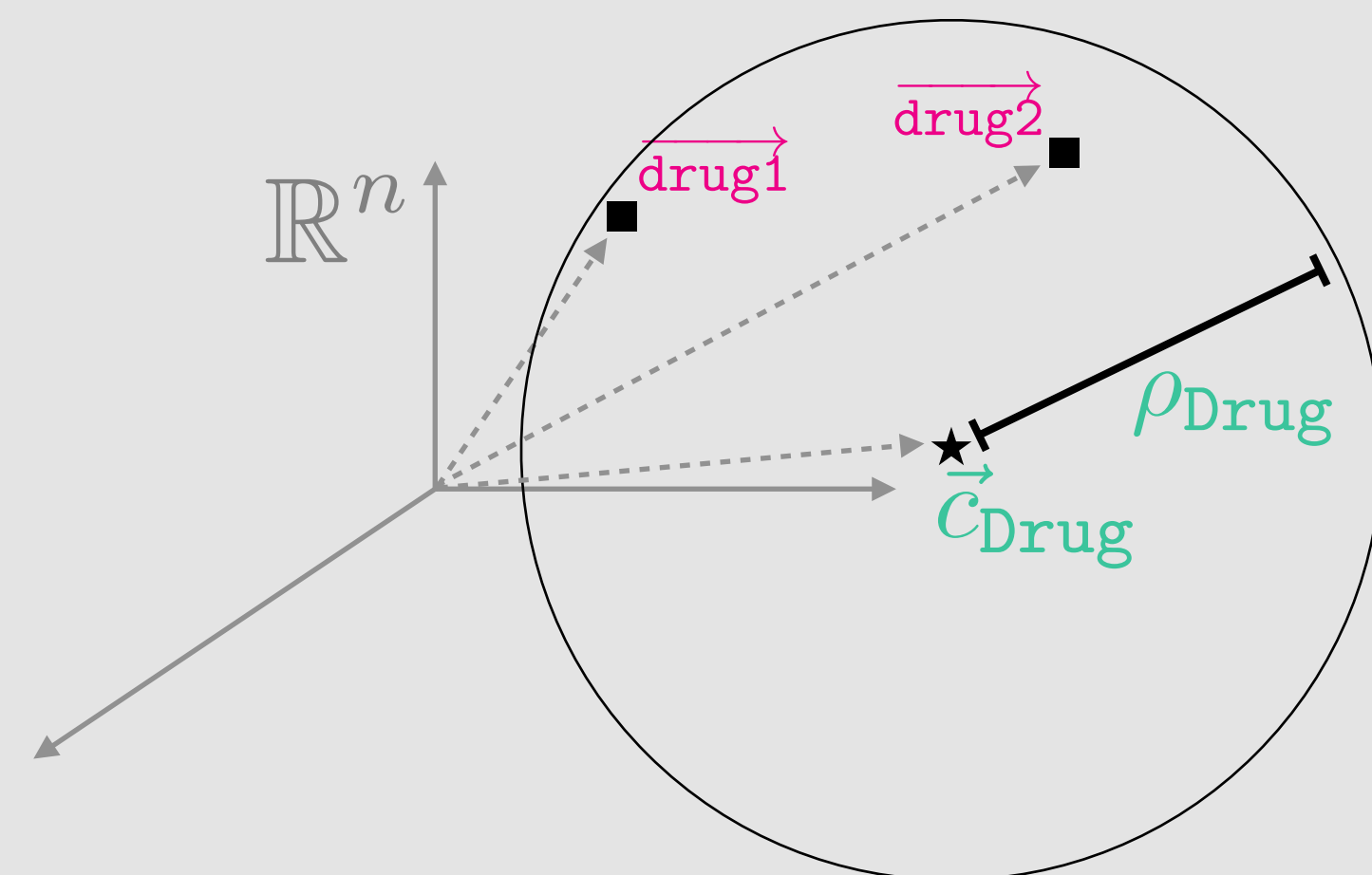
This can result in inferred triples (i.e. facts) that violate the semantic constraints of the graph:

(uri7, uri2, uri1) \rightarrow (drug2, target, drug1)
violates semantics!

2

EmbedS modeling

In EmbedS, we model **entities** as points, **classes** as sets of points (an n -sphere, with a central vector and a radius), and **properties** as sets of pairs of points (modeled analogously).



3

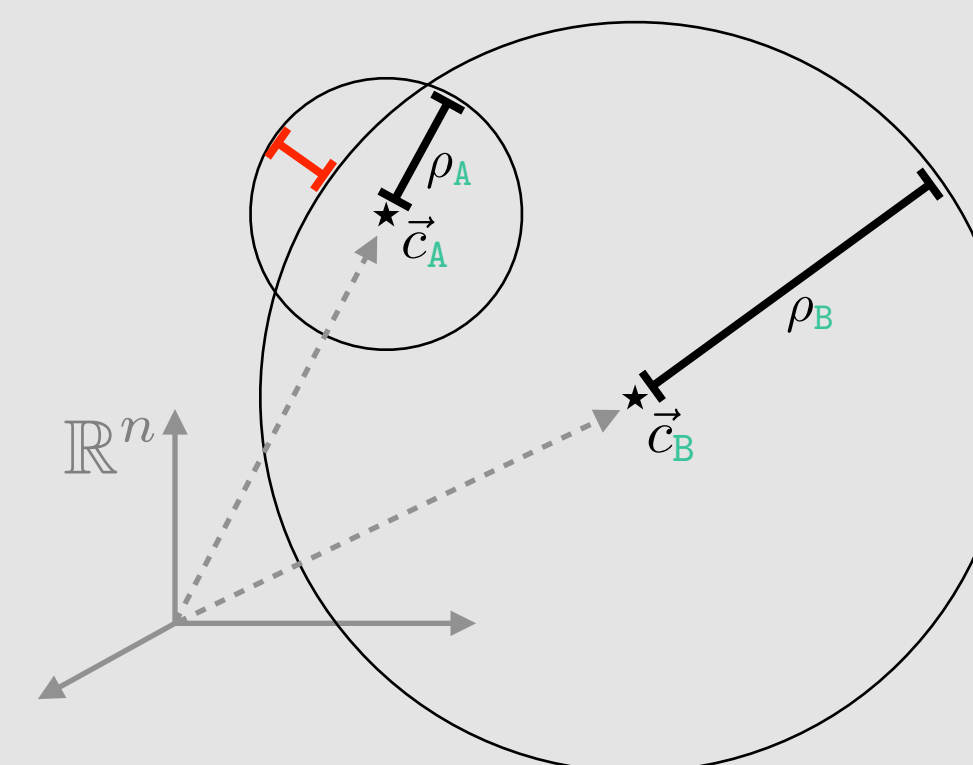
EmbedS Cost Model and Performance

EmbedS assigns a cost for violating semantic constraints. For example:

Example: for an RDFS triple

(A, rdfs:subClassOf, B)

the cost assigned is shown as a red error bar.



4

Hyper-parameter optimization: Random search [2].

Scalability: EmbedS uses Approximate Nearest Neighbor indexing for scalable learning.

Performance: Initial experimental evaluation on a benchmark dataset and an ad hoc dataset show competitive performance.

wn18 dataset:

hits@10: 94.9%, MRR: 0.560 (HMR: 1.79)

P = 84.2% and a Recall = 83.9%, f-measure: 84.0% (optimizing the geometrical interpretation)

dbpedia_v2 dataset:

EmbedS: hits@10: 22.7%, MRR: 0.133 (HMR: 7.52)

TransE: hits@10: 11.6%, MRR: 0.054 (HMR: 18.52)

References:

[1] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In NIPS'13.

[2] J. Bergstra and Y. Bengio. Random Search for Hyper-Parameter Optimization. JMLR'12.