Taming Knowledge Graphs Heterogeneity and Bias in Entity Alignment

Nikolaos Fanourakis
Supervisor: Prof. Vassilis Christophides
Co-Supervisor: Assistant Prof. Vasilis Efthymiou

FORTH
INSTITUTE OF COMPUTER SCIENCE

H.F.R.I.
Hellenic Foundation for Research & Innovation

Outline

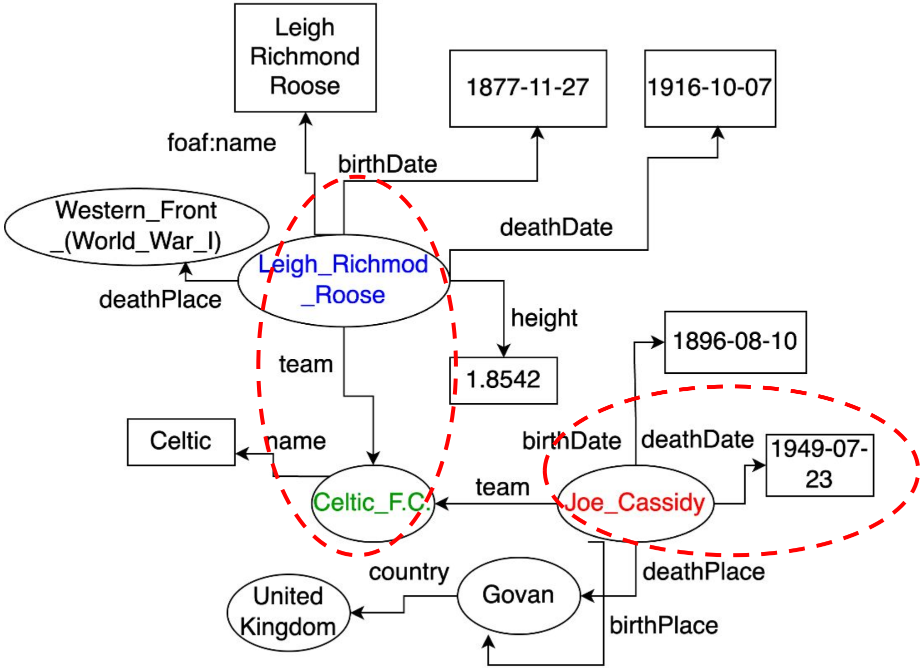
- Introduction
  - Knowledge Graphs (KG)
  - Entity Alignment (EA) Problem
  - KGs Heterogeneity & Structural Bias
  - Problem Statement & Main Challenges
  - Contributions
- Quantitative Analysis of EA Datasets
- HybEA: An Adaptable Framework for EA with KG Embeddings
- SUSIE: An Exploration-based Sampling Algorithm
  - HybEA Robustness to Structural Bias of KGs
- A Fairness-aware EA system
- Conclusions & Future Work

Knowledge Graphs (KG)

$$KG = (E, R, A, L, X, Y)$$

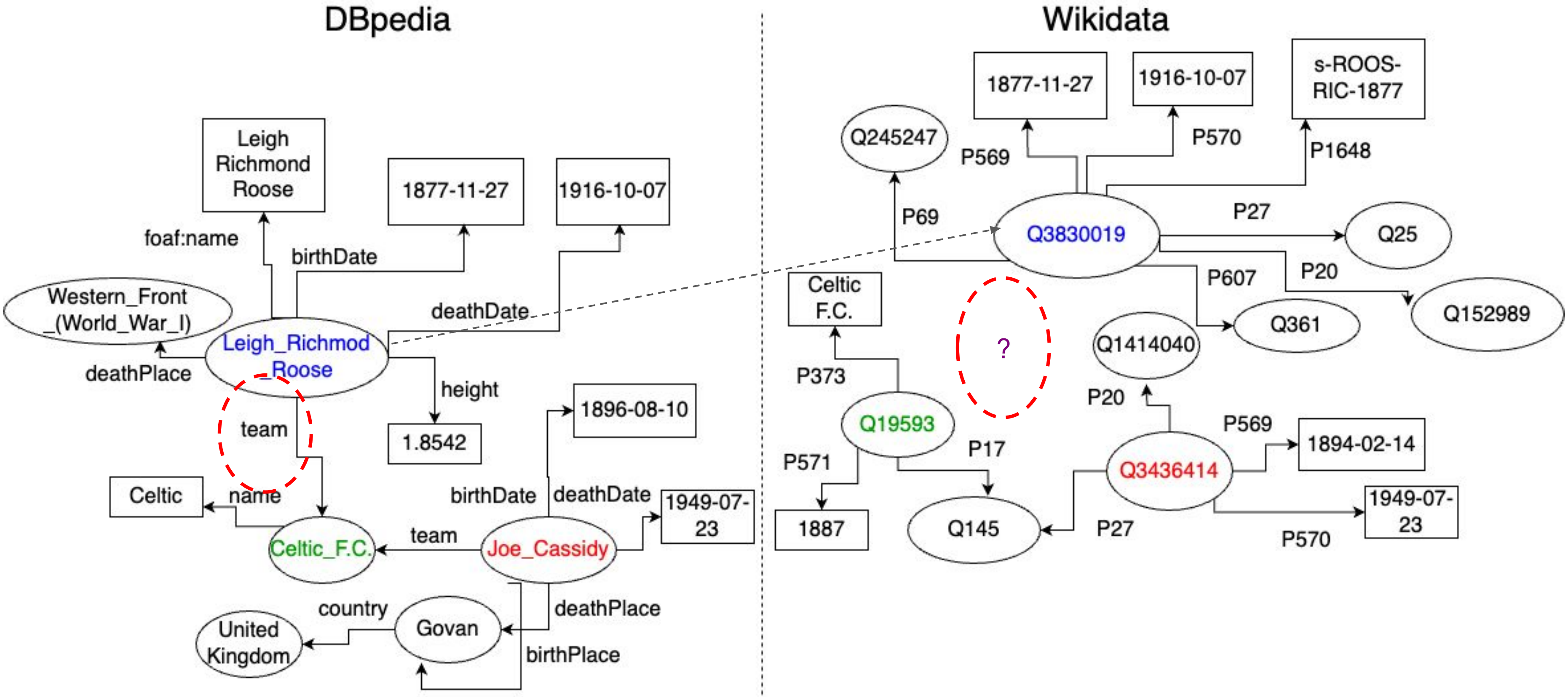
E, R, A and **L** are the sets of entity names, relation types, attribute types, and literals, respectively

DBpedia



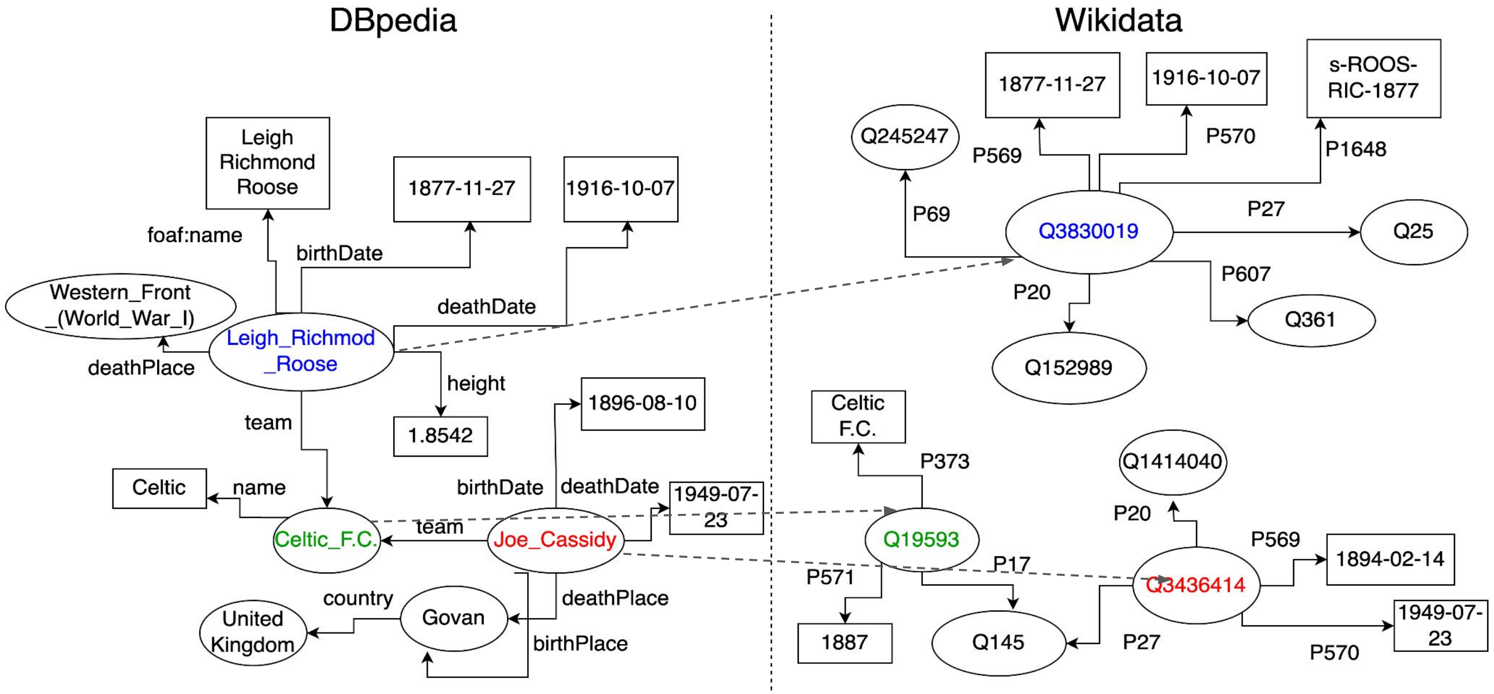
- Knowledge Graphs (e.g., DBpedia) describe entities in a machine-readable way

Integrating Entities From Multiple KGs



- Incomplete descriptions of the same real-world entities
- Increase completeness of entity descriptions

Entity Alignment (EA)



Given a source $KG_1 = (E_1, R_1, A_1, L_1, X_1, Y_1)$

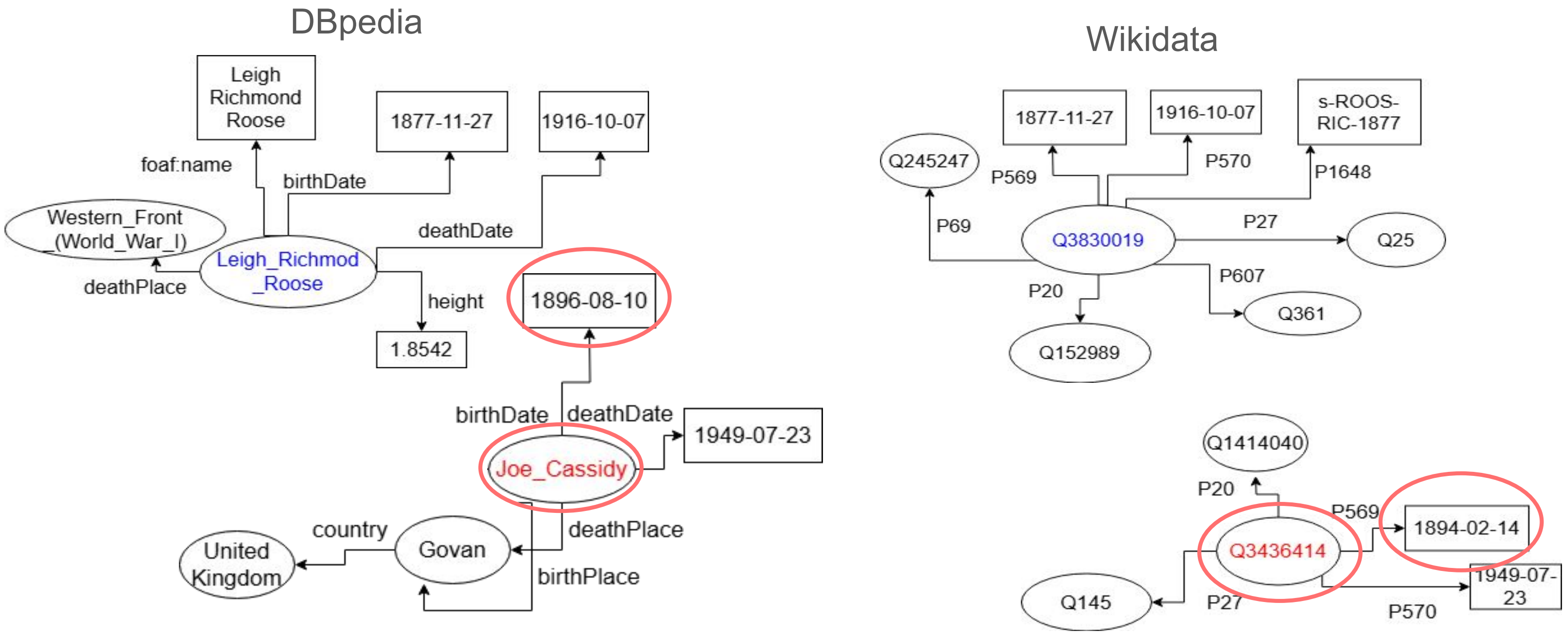
and a target $KG_2 = (E_2, R_2, A_2, L_2, X_2, Y_2)$

find pairs of matches $M = \{(e_i, e_j) \in E_1 \times E_2 \mid e_i \equiv e_j\}$

Common assumption:

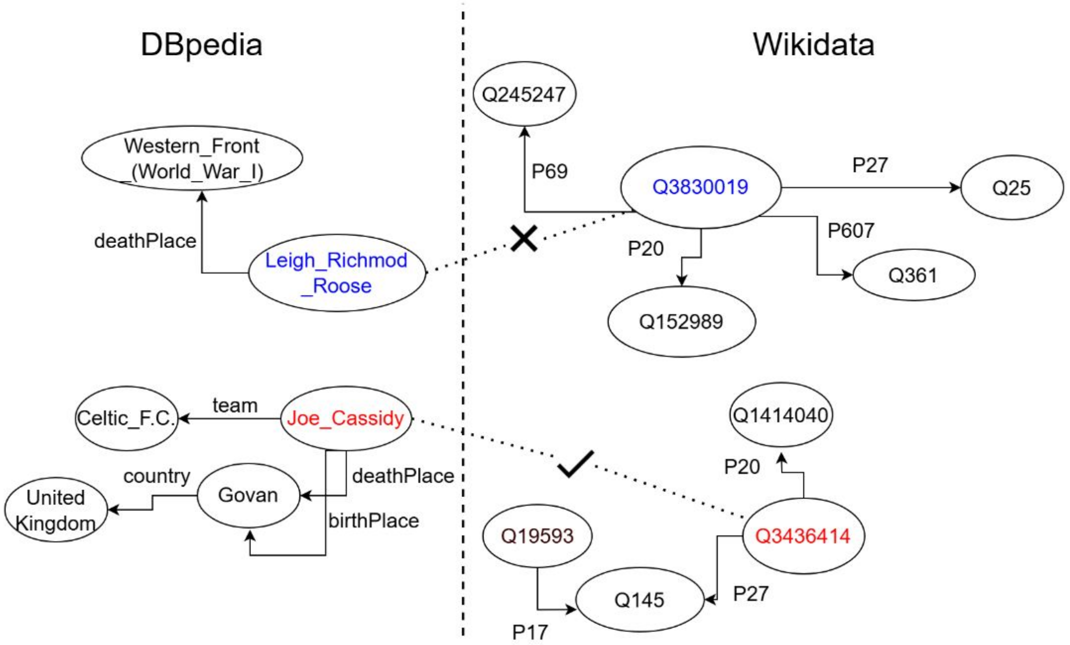
one-to-one mapping

Heterogeneity of KGs



- Structural heterogeneity of KGs
  - non-isomorphic, diverse neighborhoods
- Factual heterogeneity of KGs
  - different entity names
  - different literals
- We cannot a priori prioritize one over the other heterogeneity forms

Structural (Indirect) Bias of KGs



- **Incompletely** described entities (e.g., missing relations) lead to **structural diversity** of KGs (size and number of **connected components)**
  - **Structural bias** is a special case of **indirect bias** (sampling, representation) against protected groups defined over sensitive attributes (e.g., gender, race)
  - Exploit factual information to complete the structural similarity of entities

Problem Statement

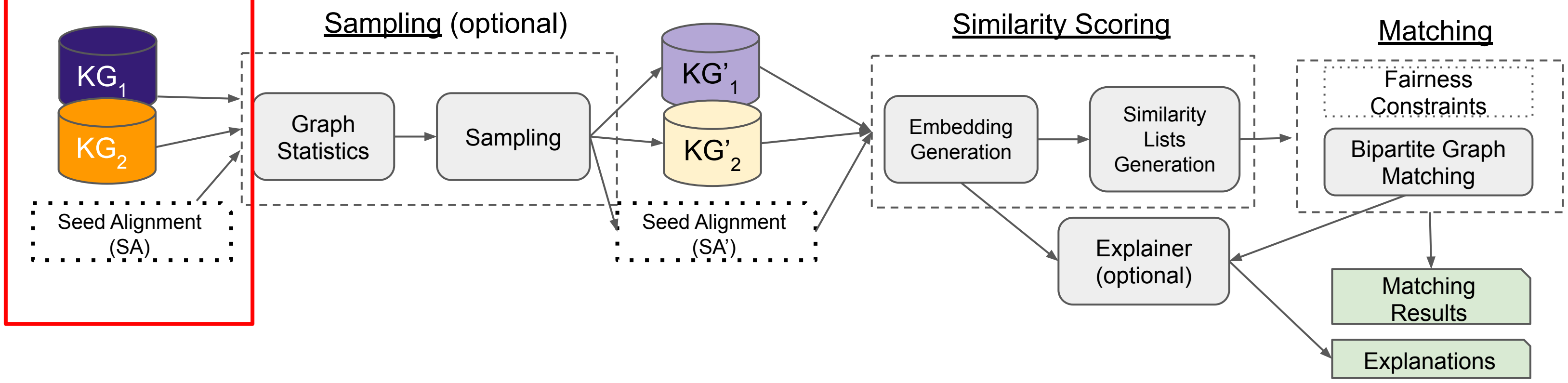
- Align entities in an **adaptive way** to different **degrees** of factual and structural **heterogeneity** exhibited by real KGs, **robust to structural bias**

- Main Challenges
    - Wide **variety of graph representation learning** to create entity embeddings based on different assumptions regarding KGs heterogeneity
    - **Labels scarcity** requires **semi-supervised** frameworks leveraging entity embeddings generated by the structural and factual information of entities.

Contributions

- New EA problem: **Quantitative analysis** of datasets with respect to **different levels of structural & factual heterogeneity**
  - 7 monolingual and 3 multilingual datasets
- New EA method: **Semi-supervised** framework for building **hybrids** of a novel **factual-based** EA model with several existing **structural-based** EA methods
  - HybEA: An adaptable framework for EA
- New EA sampling algorithm: For **assessing robustness** of the EA methods to **structural bias** of KGs
  - SUSIE: Exploration-based sampling algorithm
- New EA system: A **fairness-aware** EA system
- Thorough empirical study: Comparison with **11 baseline** methods over **10 datasets** used in previous works

1. Quantitative Analysis of EA Datasets



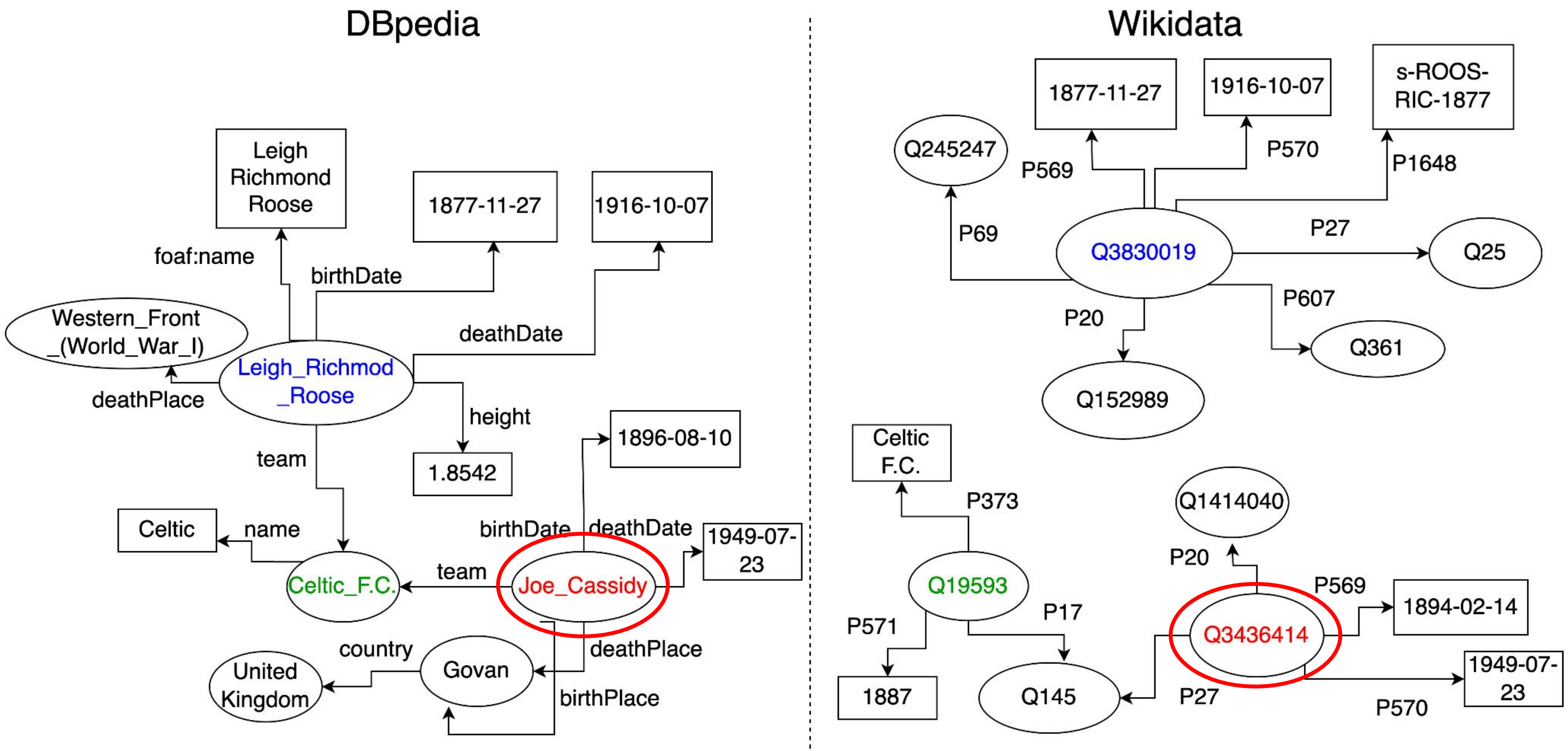
Heterogeneity Metrics

	Factual Heterogeneity	Structural Heterogeneity	Structural Bias
lev_{index} (for entity names) ↓	✓	✗	✗
lev_{index} (for attributes) ↓	✓	✗	✗
Jaccard ↓	✗	✓	✗
LDMAD ↑	✗	✓	✗
wccR ↑	✗	✓	✓
maxCS ↓	✗	✓	✓
$\overline{deg}$ ↓	✗	✓	✓

Measuring Factual Heterogeneity

$$\mathrm{lev}_{\mathrm{index}}(a, b) = \frac{|a| + |b| - \mathrm{lev}_{\mathrm{distance}}(a, b)}{|a| + |b|}$$

$\mathrm{lev}_{\mathrm{index}}$(“Joe_Cassidy”, “Q3436414”) = **0.4**

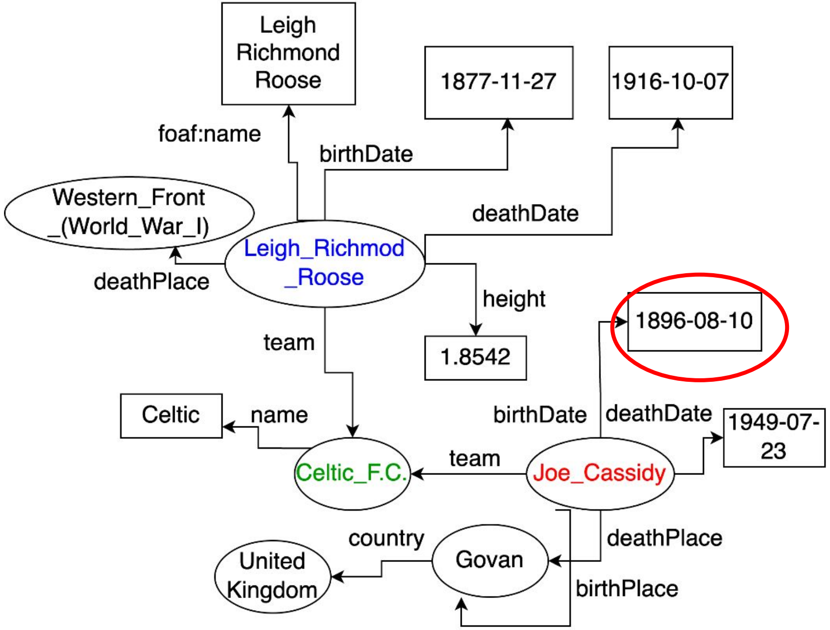


Measuring Factual Heterogeneity

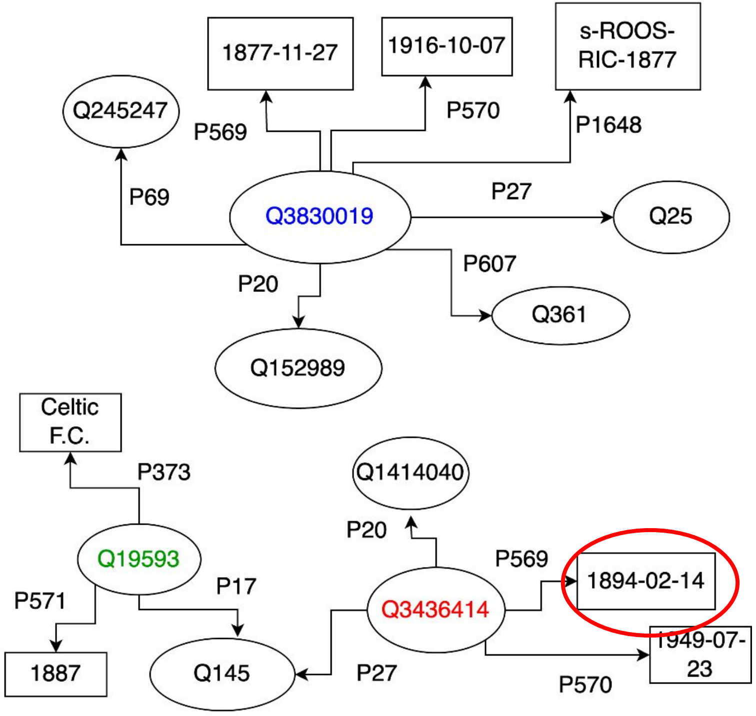
$$\mathrm{lev}_{\mathrm{index}}(a,b) = \frac{|a| + |b| - \mathrm{lev}_{\mathrm{distance}}(a,b)}{|a| + |b|}$$

$\mathrm{lev}_{\mathrm{index}}$(“1896-08-10”, “1894-02-14”) = **0.7**

DBpedia



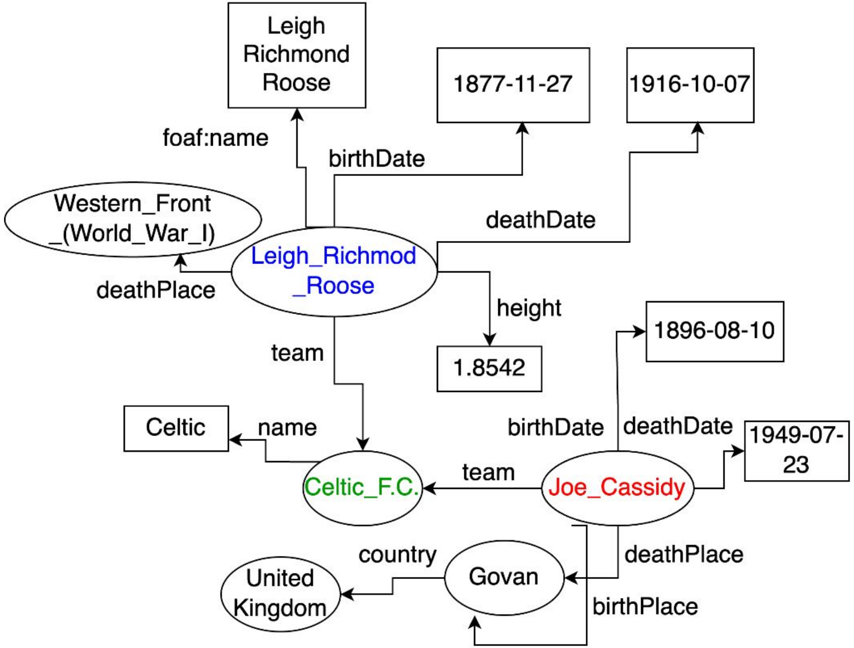
Wikidata

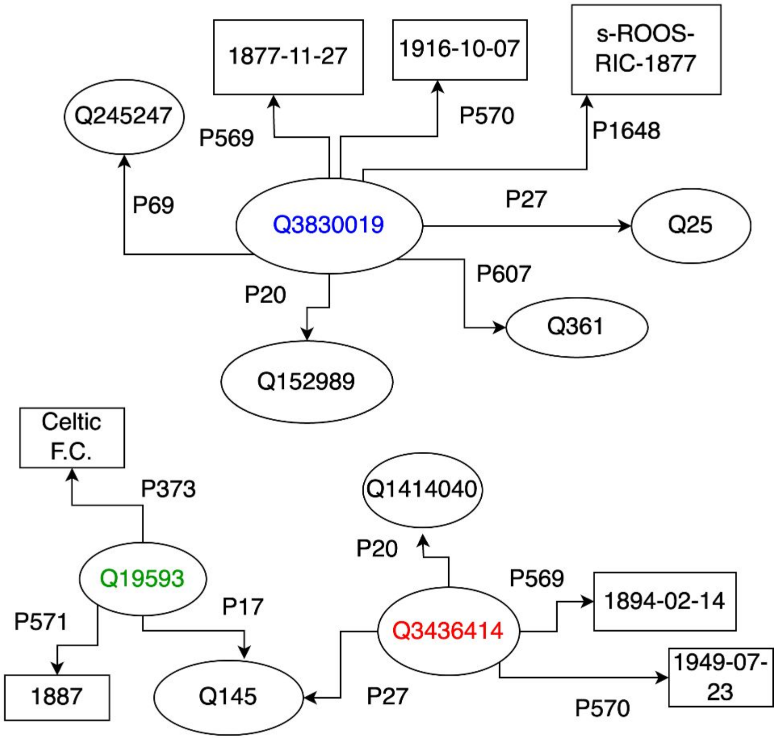


Measuring Structural Heterogeneity

Jaccard index between the matched neighbors = (1/(2 + 4 - 1) + 0 + 1/(2 + 2 - 1)) / 3 = **0.17**

$$J = \frac{1}{|M|} \sum_{(u,v) \in M} \frac{|M_{u,v}|}{|\mathcal{N}(u)| + |\mathcal{N}(v)| - |M_{u,v}|}$$

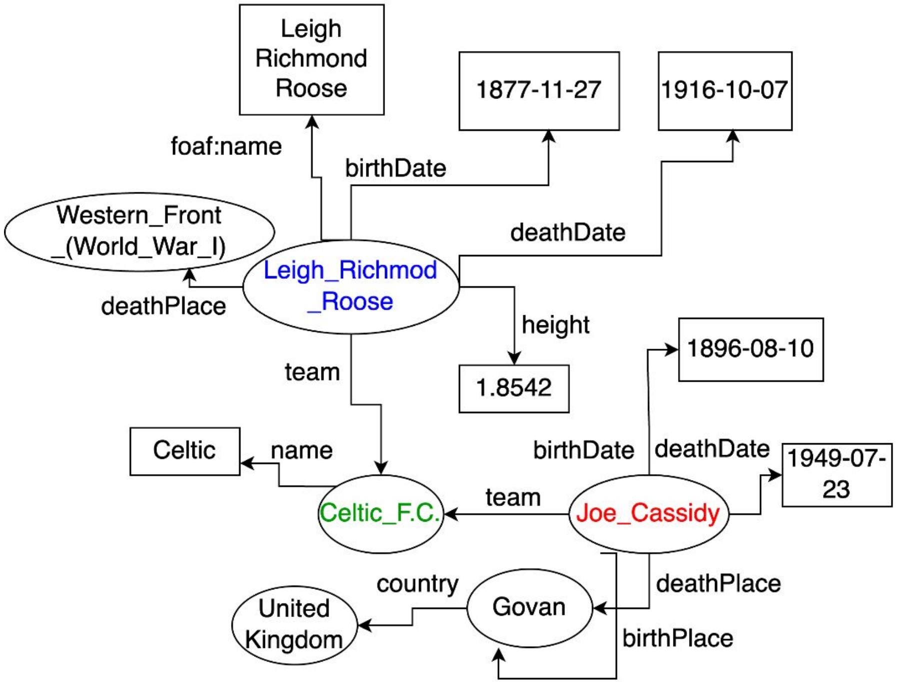


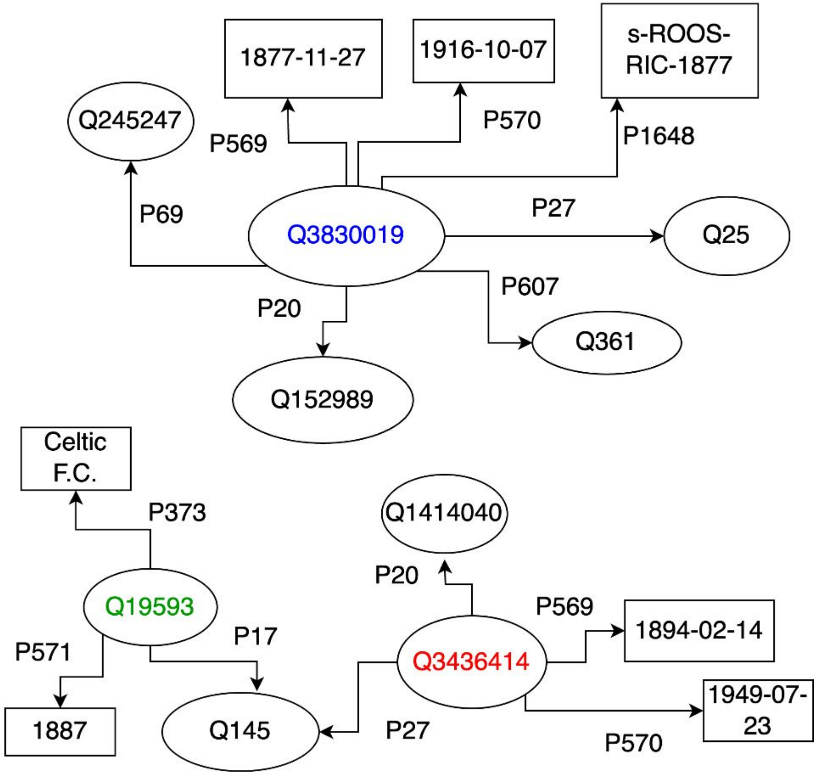


Measuring Structural Heterogeneity

Local Degree Mean Absolute Deviation = (|2 - 4| + |2 - 1| + |2 - 2|) / 3 = **1**

$$\mathrm{LDMAD} = \frac{1}{|M|} \sum_{(u,v)\in M} \left\| \mathcal{N}(u)| - |\mathcal{N}(v) \right\|$$

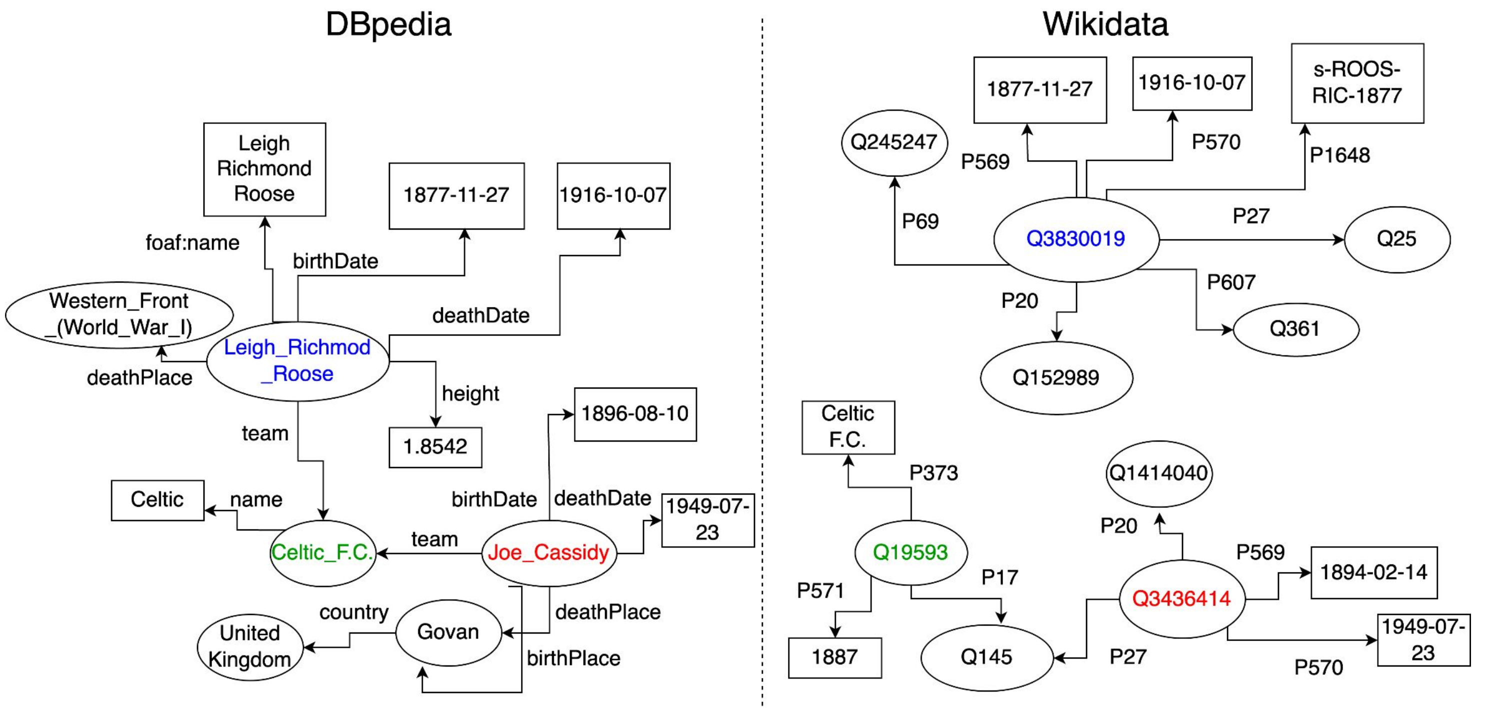




Measuring Structural Heterogeneity

Ratio of Weakly Connected Components

$$wccR(KG) = |wcc(KG)|/|E|$$

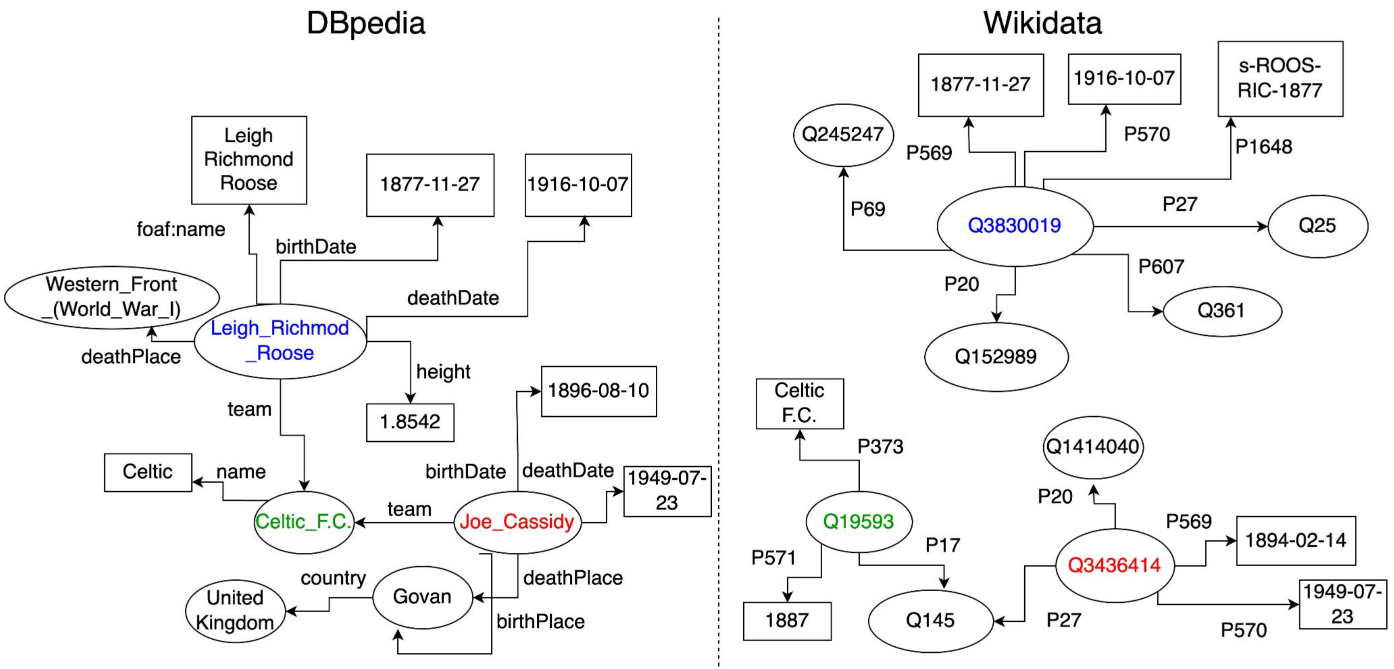


wccR = 1 / 6 = **0.16**

wccR = 2 / 9 = **0.22**

Measuring Structural Heterogeneity

Normalized Max Component Size $maxCS(KG) = \left(\max_{CC \in wcc(KG)}(|CC|)\right) / |E|$



maxCS = 6 / 6 = **1**

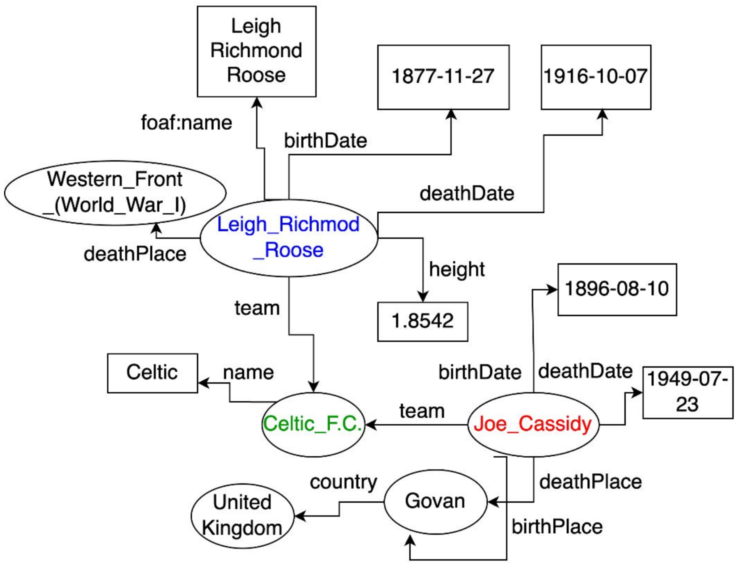
maxCS = 5 / 9 = **0.55**

Measuring Structural Heterogeneity

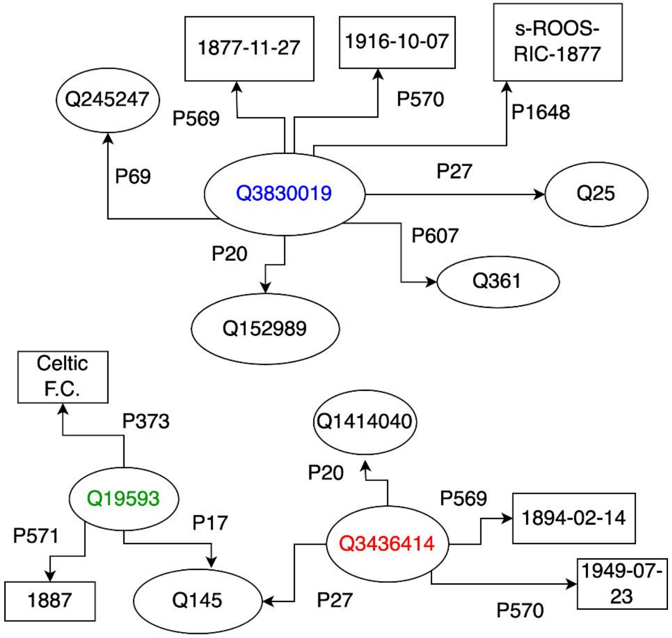
Average Node Degree

$$\overline{deg}(KG) = \frac{1}{|E|} \sum_{e_i \in E} deg\left(e_i\right)$$

DBpedia



Wikidata



$\overline{\text{deg}}$ = (1 + 2 + 2 + 2 + 1 + 3) / 6 = **1.83**

$\overline{\text{deg}}$ = (4+1+2+1+2+1+1+1+1) / 9 = **1.55**

EA Datasets With Different Degrees of Heterogeneity

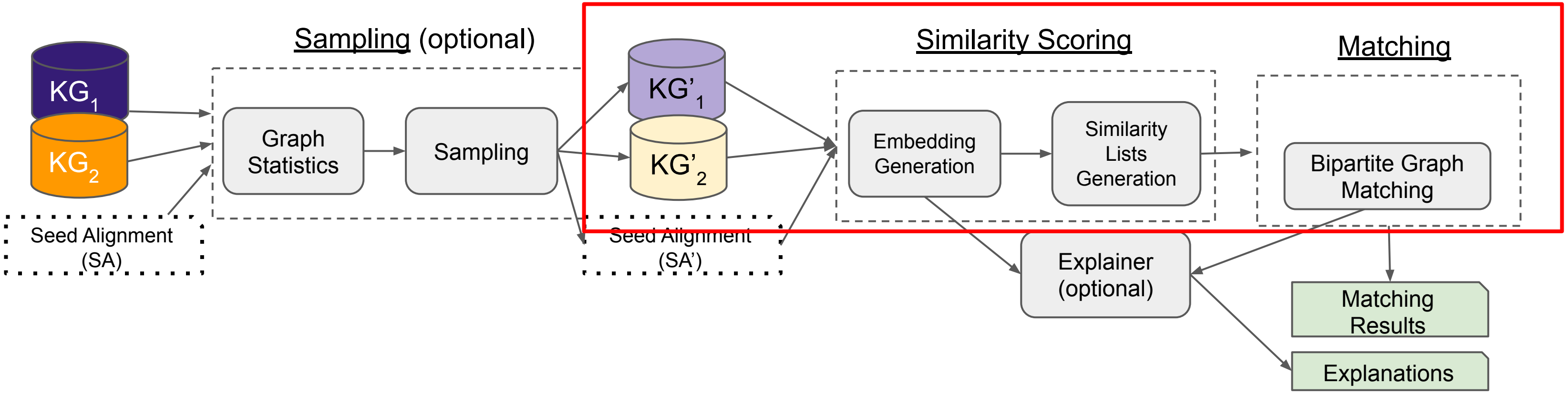
Heterogeneity	high factual	medium factual	low factual
high structural			ICEWS-WIKI, ICEWS-YAGO
medium structural		JA-EN ZH-EN	FR-EN
low structural	D-W(S), D-W(D), D-W(SRPRS-N), D-W(SRPRS-D)		BBC-DB

Simple-HHEA

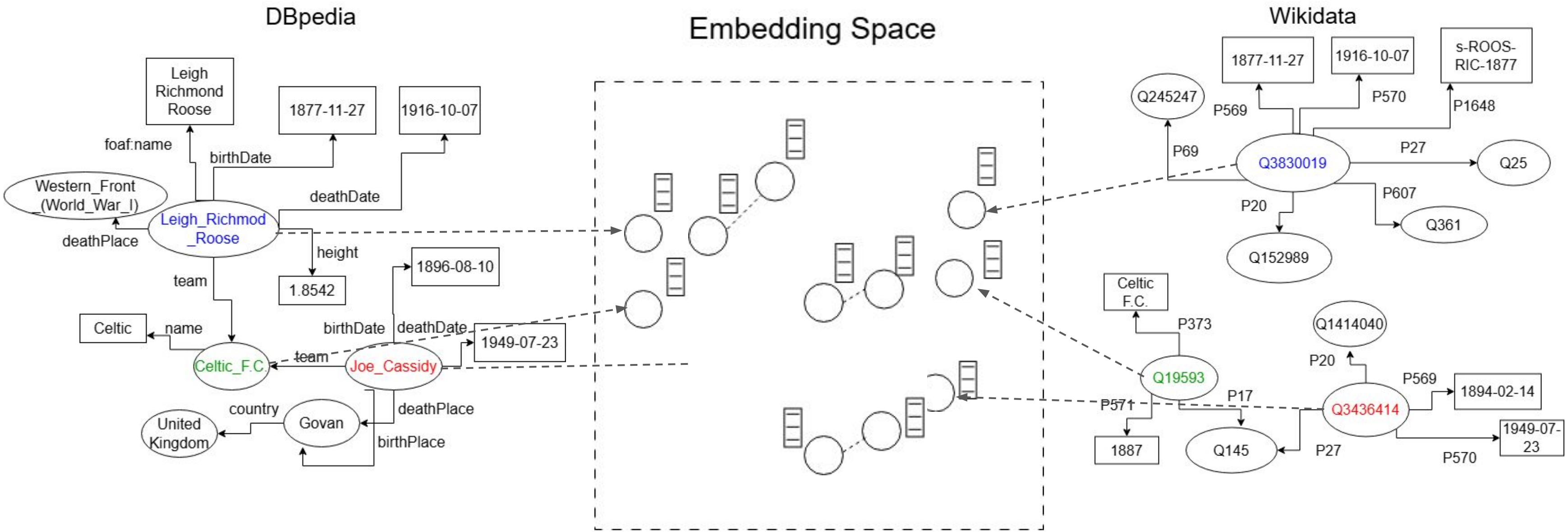
COTSAE

MTransE, Knowformer, **BERT-INT**, RREA, **PipEA**, **ZeroEA**, **SelfKG**, **AttrGNN**

2. HybEA: An adaptable framework for EA

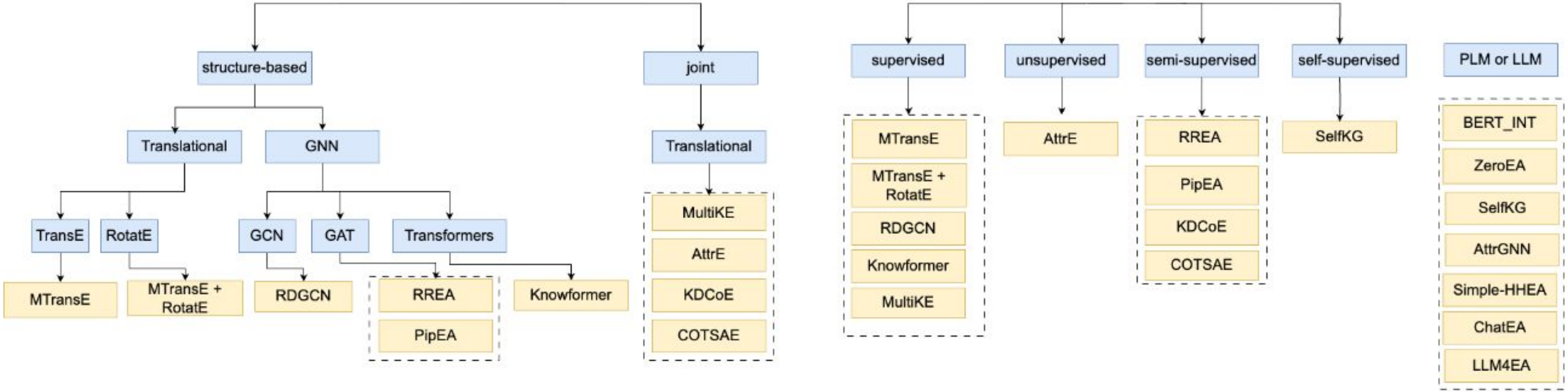


Knowledge Graph Embeddings (KGE) for EA



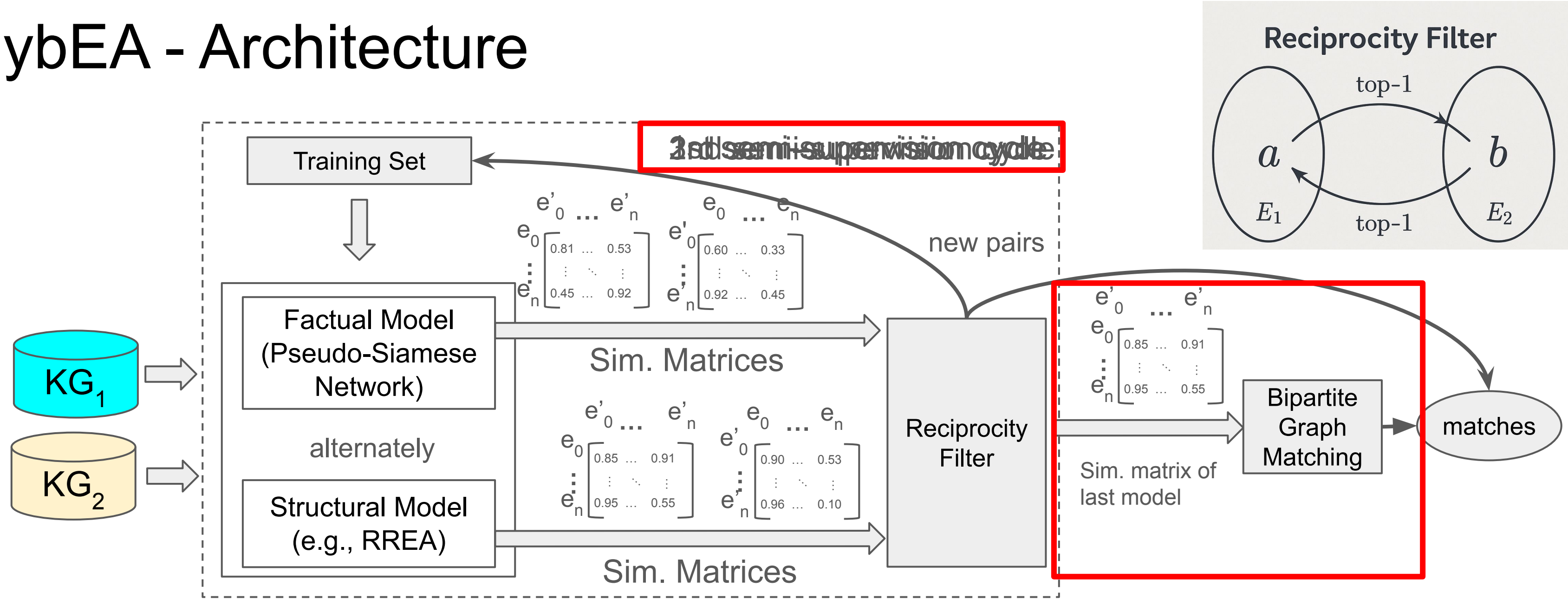
- Solve EA as a **Representation Learning** task in an embeddings space + **Similarity-based Matching**

Limitations of SOTA KGE Methods for EA



- Translational and GNN-based methods assume <u>low/medium heterogeneity</u>
- Supervised methods require large amounts of <u>labeled data</u>
- Language Models-based methods incur multiple LLM queries with a <u>important monetary cost</u>

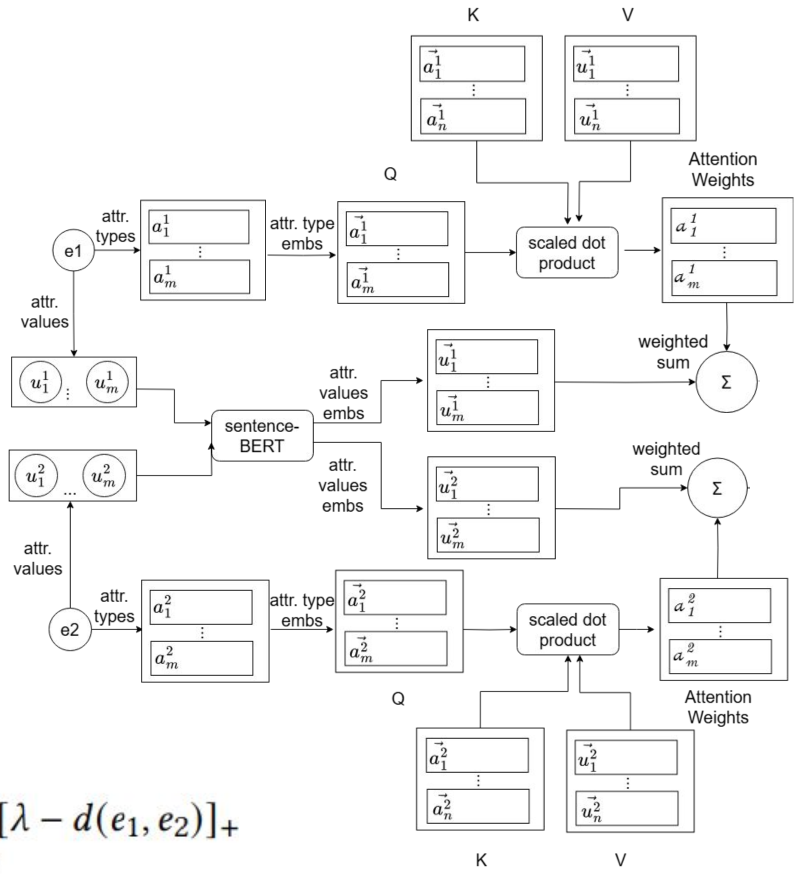
HybEA - Architecture



- **Reciprocity filter:** high-confidence matching pairs feed back to the models and they are part of framework's returned matches
  - low cost and accurate (100% precision)
- For the remaining entities, run **bipartite graph matching** on the sim. matrix of the last model

HybEA: Factual Model

- Exploit all attributes of KGs for learning the attribute attentions
- Use contextualized attribute value embeddings



Scaled dot product

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$$\vec{e} = \sum_{(e,a_i,v_i)\in Y} a_i \vec{v}_i$$

Contrastive loss

$$\mathcal{L}_a = (1-\alpha) \sum_{(e_1,e_2)\in M} d(e_1, e_2) + \alpha \sum_{(e'_1,e'_2)\in N} [\lambda - d(e_1, e_2)]_+$$

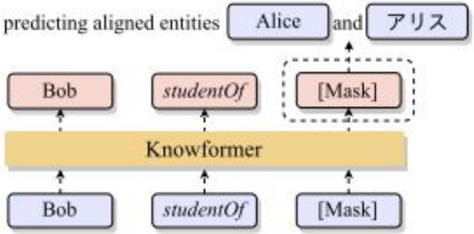
HybEA Structural Model: Knowformer

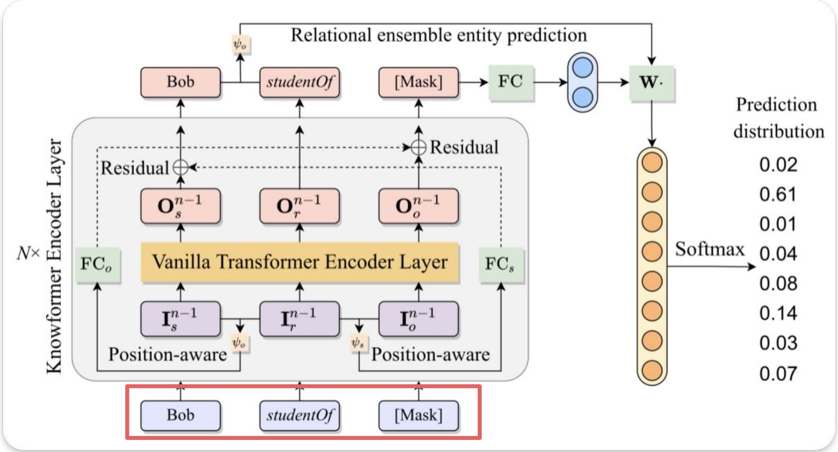
- Triple level embeddings
- Inject relational knowledge into features vector (relational composition)
- Capture positions of entities in the relational triple (ψ_s and ψ_o)
- Cross entropy loss

$$\mathrm{PAR}(s, r, o) = s + \alpha \cdot h_s(\psi_s(r, o))r, o + \alpha \cdot h_o(\psi_o(r, s))$$

$$\psi_s(r, o) = o - r$$

$$\psi_o(r, s) = s + r.$$





Li et al. "Position-Aware Relational Transformer for Knowledge Graph Embedding." *IEEE Transactions on Neural Networks and Learning*(2023).

HybEA Structural Model: RREA

$$L = \sum_{(e_i, e_j) \in P} \max\left(\text{dist}\left(e_i, e_j\right) - \text{dist}\left(e_i', e_j'\right) + \lambda, 0\right)$$

negative sampling (pseudo matching entity pairs)

Optimization objectives:

- minimize the distance of matched entities of the two KGs
- maximize the distance of pseudo matching entities of the two KGs

$$\boldsymbol{h}_{e_i}^{l+1} = \text{ReLU}\left(\sum_{e_j \in N_{e_i}^{e}} \sum_{r_k \in R_{ij}} \alpha_{ijk}^{l} \boldsymbol{M}_{r_k} \boldsymbol{h}_{e_j}^{l}\right)$$

embedding of e_i

neighbors embedding

graph attention

reflection transformation matrix

neighbors of e_i (wider range)

X.Mao, et al. "Relational Reflection Entity Alignment" CIKM (2020)

Baseline EA Methods

Methods	Embedding Module	Learning	Joint (structure &facts)	Entity Embeddings Initialization
MTransE	TransE	supervised	no	random
Knowformer	Transformers	supervised	no	random
BERT-INT	BERT	unsupervised	yes	entity names
RREA	GAT	semi-supervised	no	random
COTSAE	TransE	semi-supervised	yes	random
PipEA (basic)	GAT	supervised	no	random
ZeroEA	BERT	unsupervised	yes	entity names
SelfKG	GAT	self-supervised	no	entity names
Simple-HHEA	Random Walks	supervised	yes	entity names
AttrGNN	GCN	supervised	yes	entity names
PARIS+	-	rule-based	no	-
HybEA-K	**Transformers**	**semi-supervised**	**yes**	**entity names**
HybEA-R	**GAT**	**semi-supervised**	**no**	**random**

HybEA Performance (Monolingual)

H@1 Improvement

- **HybEA-R +17.6%** & **+6.6%** compared to **PARIS+**
- **HybEA-R +51%** & **+38%** compared to **RREA(basic)** and **RREA**
- **HybEA-K +64%** compared to **Knowformer**
- **HybEA-R +37%** higher than **COTSAE**

- **HybEA-K +2.8%** compared to ZeroEA (LLM-based)
- **HybEA-K +112%** compared to RREA (structure-based)

Method	Metric	D-W (S)	D-W (D)	D-W (SRPRS-N)	D-W (SRPRS-D)	BBC-DB
MTransE	H@1	0.260	0.262	0.210	0.347	0.249
	H@10	0.540	0.574	0.493	0.680	0.502
	MRR	0.35	0.36	0.30	0.46	0.33
Knowformer	H@1	0.559	0.840	0.388	0.788	0.289
	H@10	0.786	0.941	0.656	0.924	0.506
	MRR	0.64	0.87	0.47	0.83	0.37
BERT-INT	H@1	0.440	0.426	0.519	0.642	0.925
	H@10	0.489	0.485	0.534	0.650	0.937
	MRR	0.45	0.44	0.52	0.64	0.93
RREA (basic)	H@1	0.655	0.878	0.446	0.817	0.436
	H@10	0.884	0.986	0.754	0.962	0.652
	MRR	0.74	0.92	0.55	0.87	0.52
RREA	H@1	0.718	0.937	0.503	0.881	0.468
	H@10	0.900	0.991	0.768	0.977	0.651
	MRR	0.79	0.96	0.59	0.91	0.54
COTSAE	Hits@1	-	-	0.709	0.922	-
	H@10	-	-	0.904	0.983	-
	MRR	-	-	0.77	0.94	-
PipEA (basic)	H@1	0.402	0.736	0.241	0.708	0.140
	H@10	0.544	0.804	0.371	0.823	0.493
	MRR	0.45	0.76	0.29	0.75	0.02
ZeroEA	H@1	0.466	0.538	0.469	0.604	0.969
	H@10	0.504	0.567	0.488	0.620	0.976
	MRR	0.43	0.48	0.45	0.55	0.80
SelfKG	H@1	0.542	0.620	0.586	0.734	0.961
	H@10	0.707	0.749	0.750	0.852	0.989
	MRR	0.59	0.66	0.63	0.77	0.97
Simple-HHEA	H@1	0.071	0.077	0.104	0.144	0.098
	H@10	0.235	0.327	0.305	0.402	0.291
	MRR	0.12	0.15	0.17	0.22	0.16
AttrGNN	H@1	0.522	0.570	0.366	0.193	0.311
	H@10	0.692	0.657	0.588	0.343	0.539
	MRR	0.58	0.60	0.44	0.24	0.39
PARIS+	H@1	0.841	0.938	0.442	0.834	0.387
HybEA-R	H@1	**0.989**	**1.000**	**0.972**	**0.997**	0.993
	H@10	**0.997**	**1.000**	**0.992**	**1.000**	**1.000**
	MRR	**0.99**	**1.00**	**0.98**	**1.00**	**1.00**
HybEA-K	H@1	0.920	0.988	0.908	0.968	**0.996**
	H@10	0.969	0.997	0.954	0.987	0.997
	MRR	0.93	0.99	0.92	0.97	0.99
Δ	H@1	+0.148 (+17.6%)	+0.062 (+6.6%)	+0.263 (+40.6%)	+0.075 (+7.7%)	+0.027 (+3.6%)

HybEA Performance (Monolingual)

- HybEA on average **17.95%** improvement compared to best-performing baseline

H@1 Improvement

- HybEA-R: +**1888%** and +**6112%** compared to RREA(basic) and Knowformer (structure-based)
- HybEA-R: +**18.5%** compared to SelfKG (LLM-based)
- HybEA-R: +**38%** compared to Simple-HHEA

Method	Metric	ICEWS-WIKI	ICEWS-YAGO
MTransE	H@1	0.001	0.000
	H@10	0.006	0.001
	MRR	0.004	0.001
Knowformer	H@1	0.016	0.013
	H@10	0.075	0.046
	MRR	0.03	0.02
BERT-INT	H@1	0.561	0.756
	H@10	0.700	0.859
	MRR	0.60	0.79
RREA (basic)	H@1	0.050	0.026
	H@10	0.227	0.136
	MRR	0.11	0.06
RREA	H@1	0.050	0.027
	H@10	0.230	0.142
	MRR	0.11	0.06
COTSAE	Hits@1	-	-
	H@10	-	-
	MRR	-	-
PipEA (basic)	H@1	N/A	N/A
	H@10	N/A	N/A
	MRR	N/A	N/A
ZeroEA	H@1	N/A	N/A
	H@10	N/A	N/A
	MRR	N/A	N/A
SelfKG	H@1	0.839	0.806
	H@10	0.931	0.867
	MRR	0.871	0.828
Simple-HHEA	H@1	0.720	0.847
	H@10	0.872	0.915
	MRR	0.754	0.870
AttrGNN	H@1	0.047	0.015
	H@10	-	-
	MRR	0.09	0.04
PARIS+	H@1	0.672	0.687
HybEA-R	H@1	**0.994**	**0.994**
	H@10	**0.997**	**0.997**
	MRR	**1.00**	**1.00**
HybEA-K	H@1	0.916	0.941
	H@10	0.938	0.944
	MRR	0.92	0.94
Δ	H@1	+0.155 (+18.5%)	+0.147 (+17.4%)

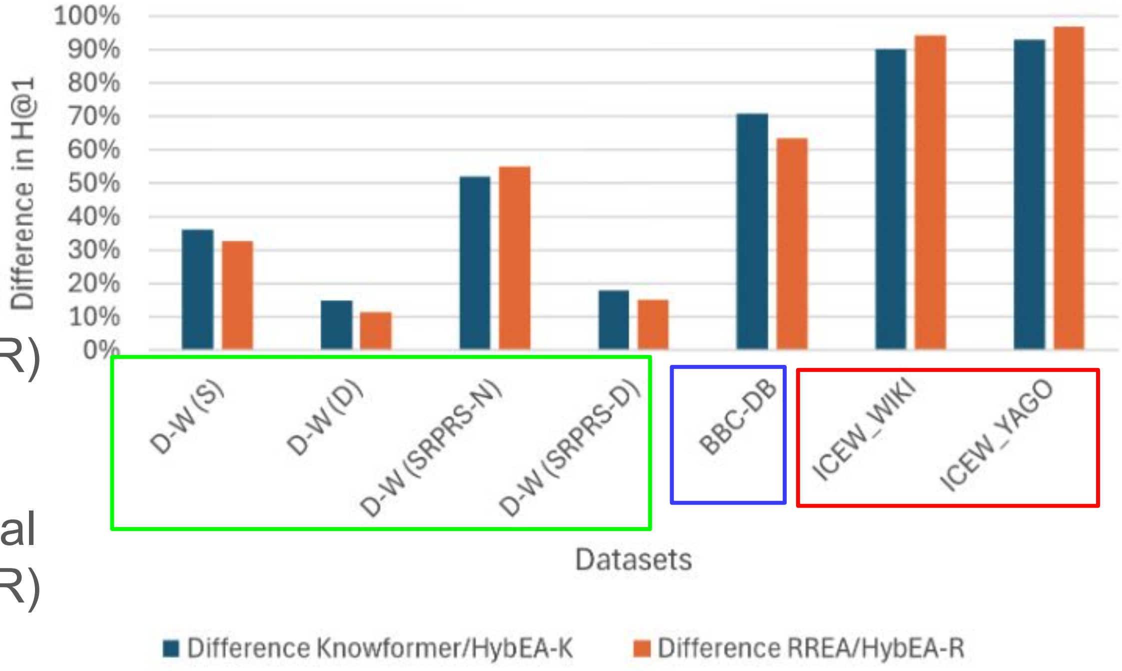
HybEA Performance (Multilingual)

- **HybEA-R** has at most **0.7%** H@1 improvement **(BERT-INT)**
- **HybEA-R** has at most **3%** H@1 improvement **(BERT-INT)**

Method	Metric	FR-EN	JA-EN	ZH-EN
MTransE	H@1	0.244	0.279	0.308
	H@10	0.556	0.575	0.614
	MRR	0.335	0.349	0.364
Knowformer	H@1	0.774	0.731	0.765
	H@10	0.932	0.902	0.888
	MRR	0.832	0.793	0.811
BERT-INT	H@1	0.992	0.964	0.968
	H@10	0.998	0.991	0.990
	MRR	0.995	0.975	0.977
RREA	H@1	0.827	0.802	0.801
	H@10	0.966	0.952	0.948
	MRR	0.881	0.858	0.857
ZeroEA	H@1	0.998	0.982	0.985
	H@10	0.999	0.995	0.993
	MRR	0.998	0.989	0.991
SelfKG	H@1	0.957	0.813	0.742
	H@10	0.992	0.906	0.861
	MRR	0.971	0.844	0.782
AttrGNN	H@1	0.942	0.783	0.796
	H@10	0.986	0.920	0.929
	MRR	0.959	0.834	0.845
PARIS+	H@1	0.882	0.824	OOM
HybEA-R	H@1	**0.999**	**0.993**	**0.994**
	H@10	**1.000**	**0.999**	**0.999**
	MRR	**0.999**	**0.995**	**0.996**

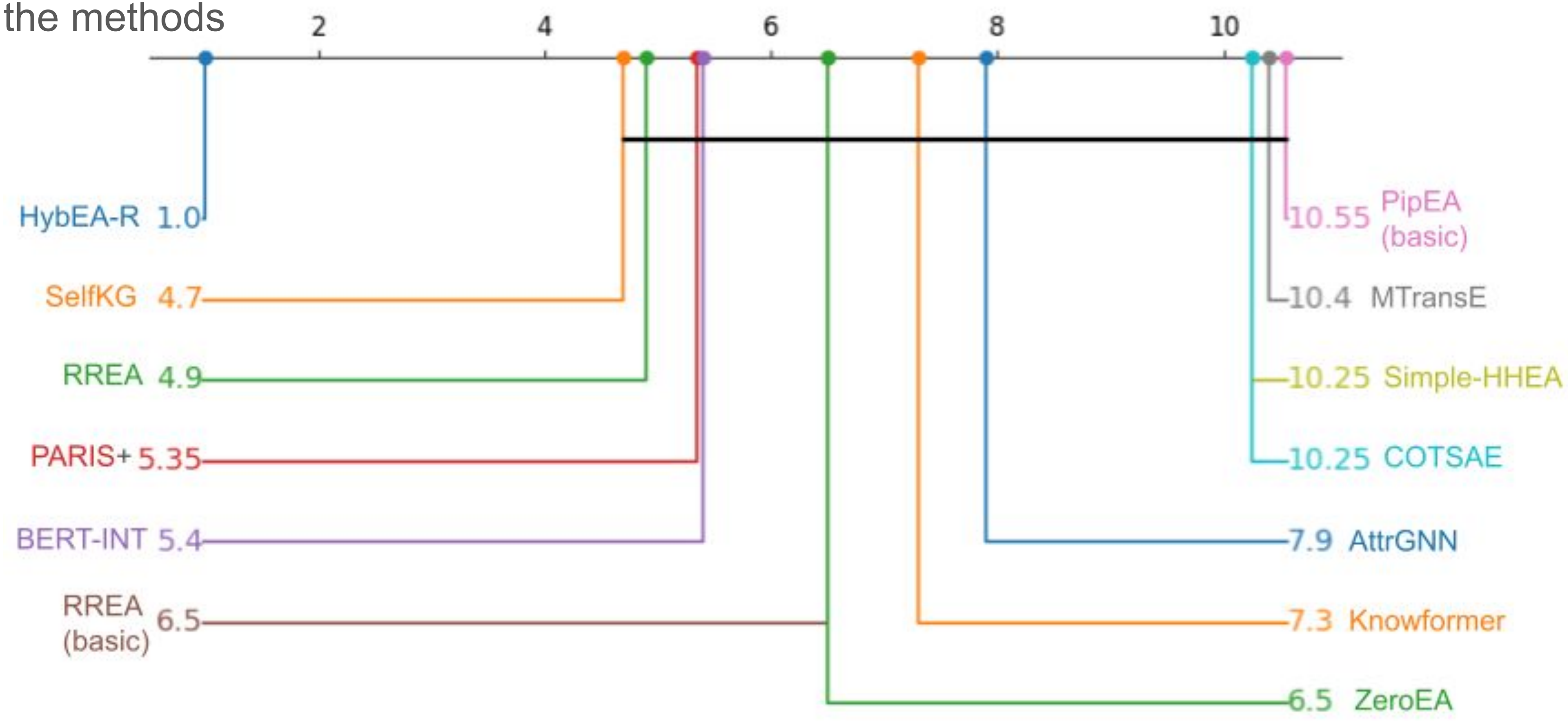
HybEA Contribution to Different Structural Models

- Contribution of HybEA in datasets:
  - Low struct. & high factual
    - 11% - 55% improv.
  - Medium struct. & factual
    - 28% improv. (HybEA-R)
  - High struct. & low factual
    - 90% - 98% improv.
  - Medium struct. & low factual
    - 25% improv. (HybEA-R)
  - Low struct. & factual
    - 64% - 70% improv.



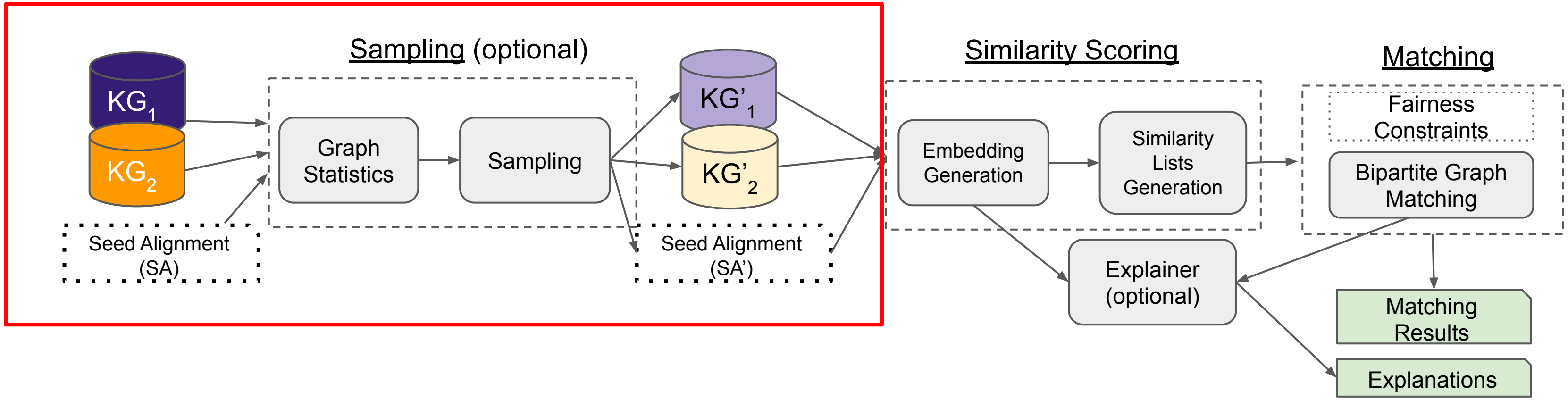
Statistical Significance Difference

Rankings of the methods



HybEA-R outperforms all baselines with statistically significant difference

3. SUSIE: An Exploration-based Sampling Algorithm



Bias in KG Modeling Tasks

		Node Classification	Recommendation	Link Prediction	Entity Alignment
Direct	Group	✓	✓	✓	✓
Indirect	Individual Fairness	✓	✗	✓	✗
	Degree-related	✓	✓	✓	✓
	Connectivity-related	✗	✓	✓	✗

Lack of publicly available benchmark data for assessing fairness of EA

Choudhary, M., Laclau, C., Largeron, C.: A survey on fairness for machine learning on graphs. CoRR abs/2205.05396 (2022)
Dong, Y., Ma, J., Chen, C., Li, J.: Fairness in graph mining: A survey. CoRR abs/2204.09888 (2022)

SUSIE

- **Random walks** on both input KGs

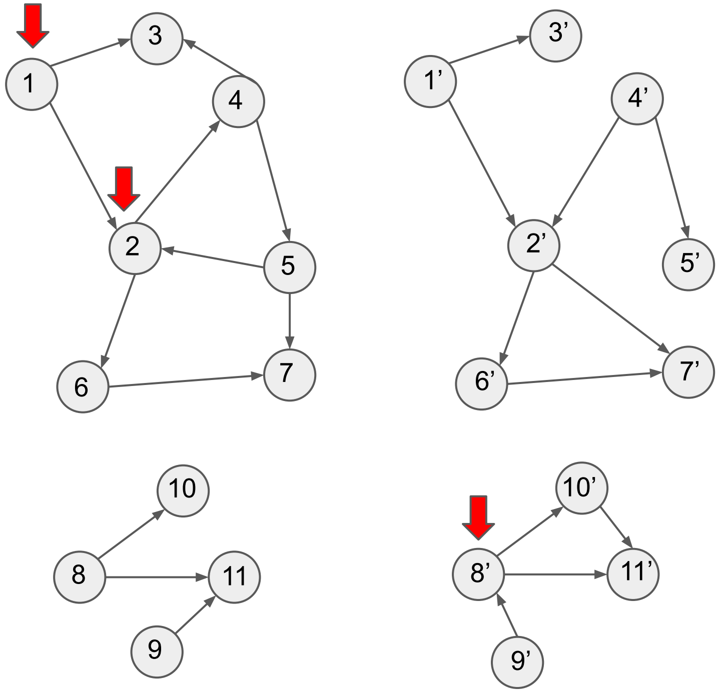
- **Jumps**, controlled by jump probability *p* (hyper-parameter):
  - visit a random, unvisited node of the other KG (swap KGs)
    - the target node belongs to a component of a randomly selected component size

- **Explores diverse areas of both KGs**, wrt the size of the connected components
  - High jump probability *p* → high structural diversity (many and small connected components in sample)

SUSIE - Methodology

Input KG_1 with 11 nodes

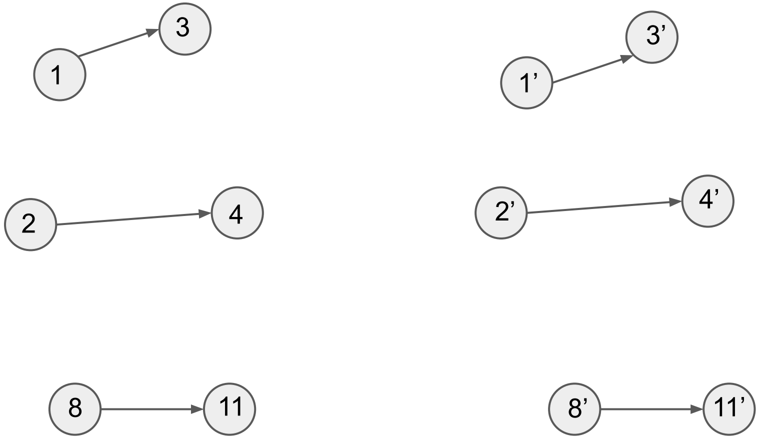
Input KG_2 with 11 nodes



- Random Walks with jump probability (*p*)
- Sampling size *s* = *6*
- *p = {0, 0.5,* ***1****}*

Sampled KG'_1 with 6 nodes

Sampled KG'_2 with 6 nodes



KGs With Different Degrees of Structural Diversity

- DBpedia - Yago (D-Y)
- DBpedia - Wikidata (D-W)
- BBC - DBpedia (BBC-DB)

	Entities ($\lvert E_1\rvert$ / $\lvert E_2\rvert$)	**Relations** ($\lvert R_1\rvert$ / $\lvert R_2\rvert$)	**Triples** ($\lvert T_1\rvert$ / $\lvert T_2\rvert$)
D-Y	15,000 /15,000	165 / 28	30,291 / 26,638
D-W	15,000 /15,000	248 / 169	38,265 / 42,746
BBC-D	9,396 / 9,396	9 / 98	15,478 / 45,561

p			input
D-Y	wccR	KG_1	0.03
		KG_2	0.04
	maxCS	KG_1	0.87
		KG_2	0.83
	$\overline{deg}$	KG_1	4.03
		KG_2	3.55
D-W	wccR	KG_1	0.01
		KG_2	0.02
	maxCS	KG_1	0.95
		KG_2	0.93
	$\overline{deg}$	KG_1	5.10
		KG_2	5.69
BBC-D	wccR	KG_1	0.18
		KG_2	0.07
	maxCS	KG_1	0.31
		KG_2	0.78
	$\overline{deg}$	KG_1	3.29
		KG_2	9.69

Baseline EA Methods

Methods	Embedding Module	Learning	Neighborhoods	Entity Embeddings Initialization
HybEA-K	Vanilla Transformers	semi-supervised	one-hop	yes
HybEA-R	GAT	semi-supervised	multi-hop	no
RREA	GAT	semi-supervised	multi-hop	no
RDGCN	GCN	supervised	multi-hop	yes
MultiKE	TransE	supervised	one-hop	no
PARIS+	-	rule-based	one-hop quasi-functional*	no

* (h,r,t) relation triples that for a given (h,r) pair, the expected number of tail entities (t) is close to 1

Controlling Structural Bias in Popular KGs

- **higher** wccR (more jumps) → less connected KG
  - **many, small** connected components
- **lower** wccR (fewer jumps) → **bigger KG regions** connected
  - **few, small connected components**
- **more** connected components→ **lower** maxCS
- **low** $\overline{deg}$(KG) (more jumps) → **high** structural diversity and sparser KGs

sampling size = 1000

p			input	0	0.15	0.5	0.85	1
D-Y	wccR	KG_1	0.03	0.01	0.15	0.24	0.28	0.28
		KG_2	0.04	0.05	0.18	0.24	0.28	0.29
	maxCS	KG_1	0.87	0.90	0.13	0.03	0.02	0.02
		KG_2	0.83	0.70	0.06	0.03	0.02	0.02
	$\overline{deg}$	KG_1	4.03	3.65	3.41	2.94	2.71	2.78
		KG_2	3.55	2.31	2.59	2.35	2.16	2.17
D-W	wccR	KG_1	0.01	0.01	0.14	0.24	0.28	0.29
		KG_2	0.02	0.03	0.11	0.20	0.24	0.24
	maxCS	KG_1	0.95	0.91	0.47	0.18	0.11	0.10
		KG_2	0.93	0.85	0.57	0.34	0.24	0.26
	$\overline{deg}$	KG_1	5.10	3.68	2.79	2.50	2.47	2.44
		KG_2	5.69	3.31	2.94	2.37	2.28	2.19
BBC-D	wccR	KG_1	0.18	0.16	0.23	0.29	0.34	0.36
		KG_2	0.07	0.01	0.16	0.24	0.28	0.31
	maxCS	KG_1	0.31	0.26	0.02	0.01	0.01	0.02
		KG_2	0.78	0.92	0.27	0.03	0.04	0.02
	$\overline{deg}$	KG_1	3.29	3.44	3.09	2.73	2.47	2.43
		KG_2	9.69	11.69	6.52	6.07	5.43	5.11

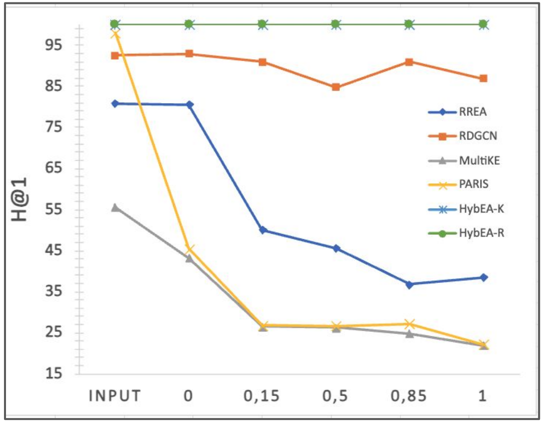
Robustness of KGE EA Methods to Structural Bias

- **HybEA-R** that also exploit factual information is the **most robust** to structural bias **in all datasets**

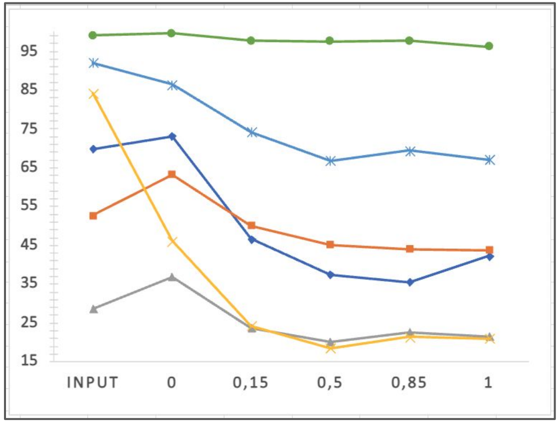
- **HybEA-K** and **RDGCN** exhibit high robustness in datasets with **low factual heterogeneity**, while in datasets with **high factual heterogeneity** less

- Conventional Ontology Alignment methods (e.g., **PARIS**) rely on largely homogeneous KGs
  - very sensitive to structural variations

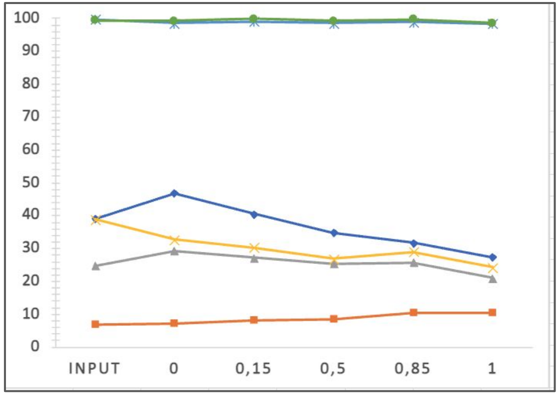
D-Y



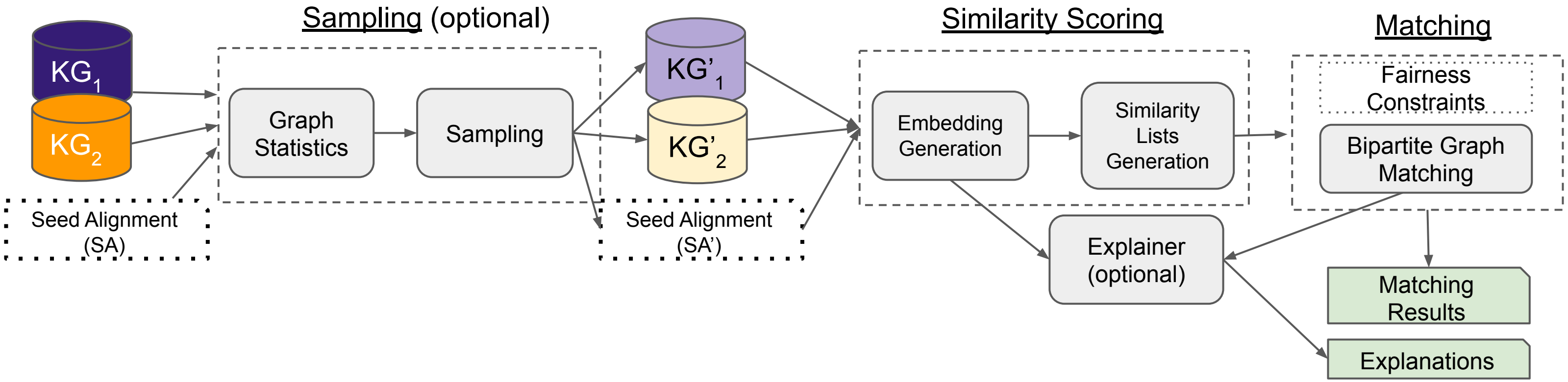
D-W



BBC-D



4. Fairness-aware EA Framework

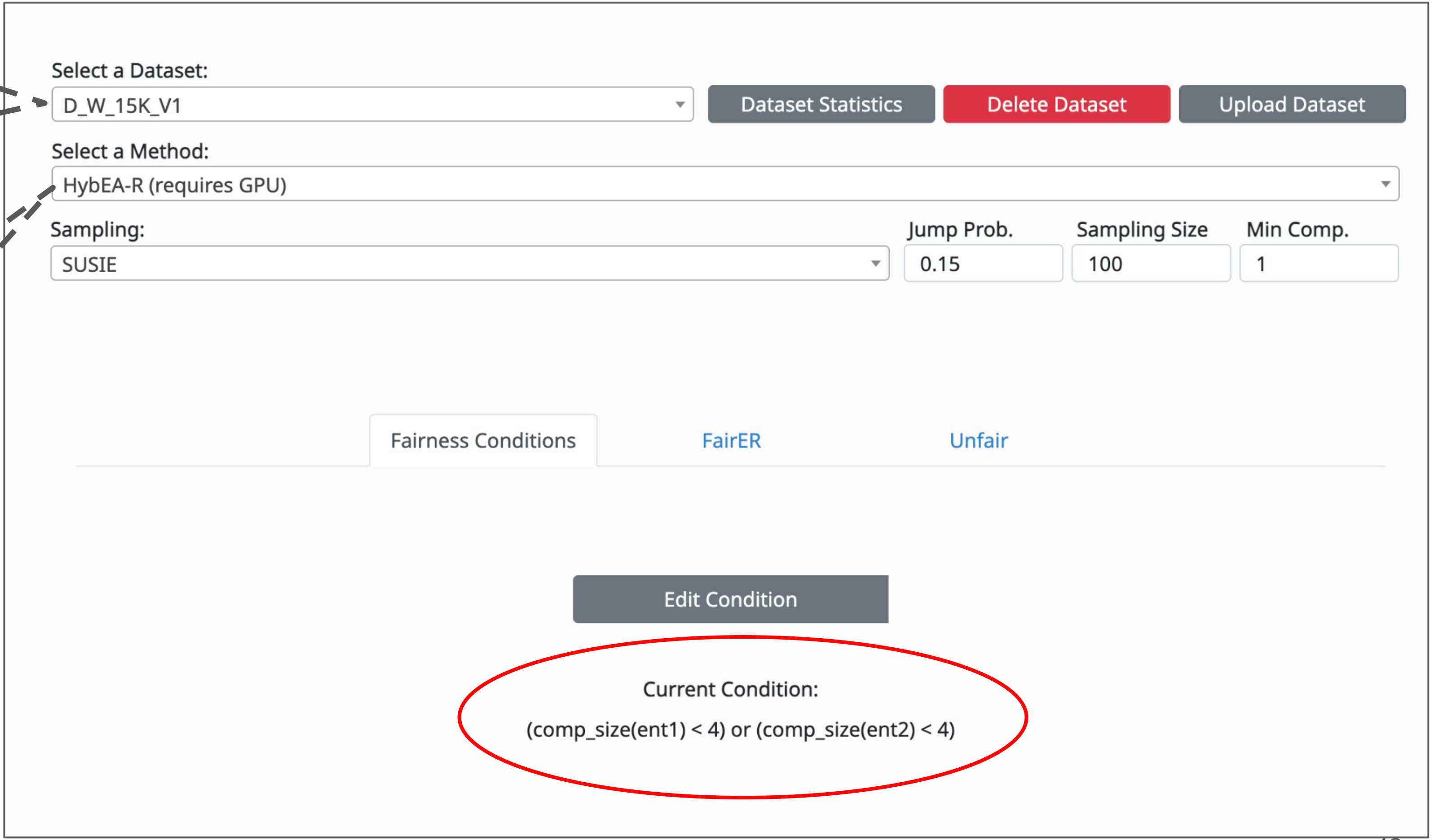


Fairness Conditions

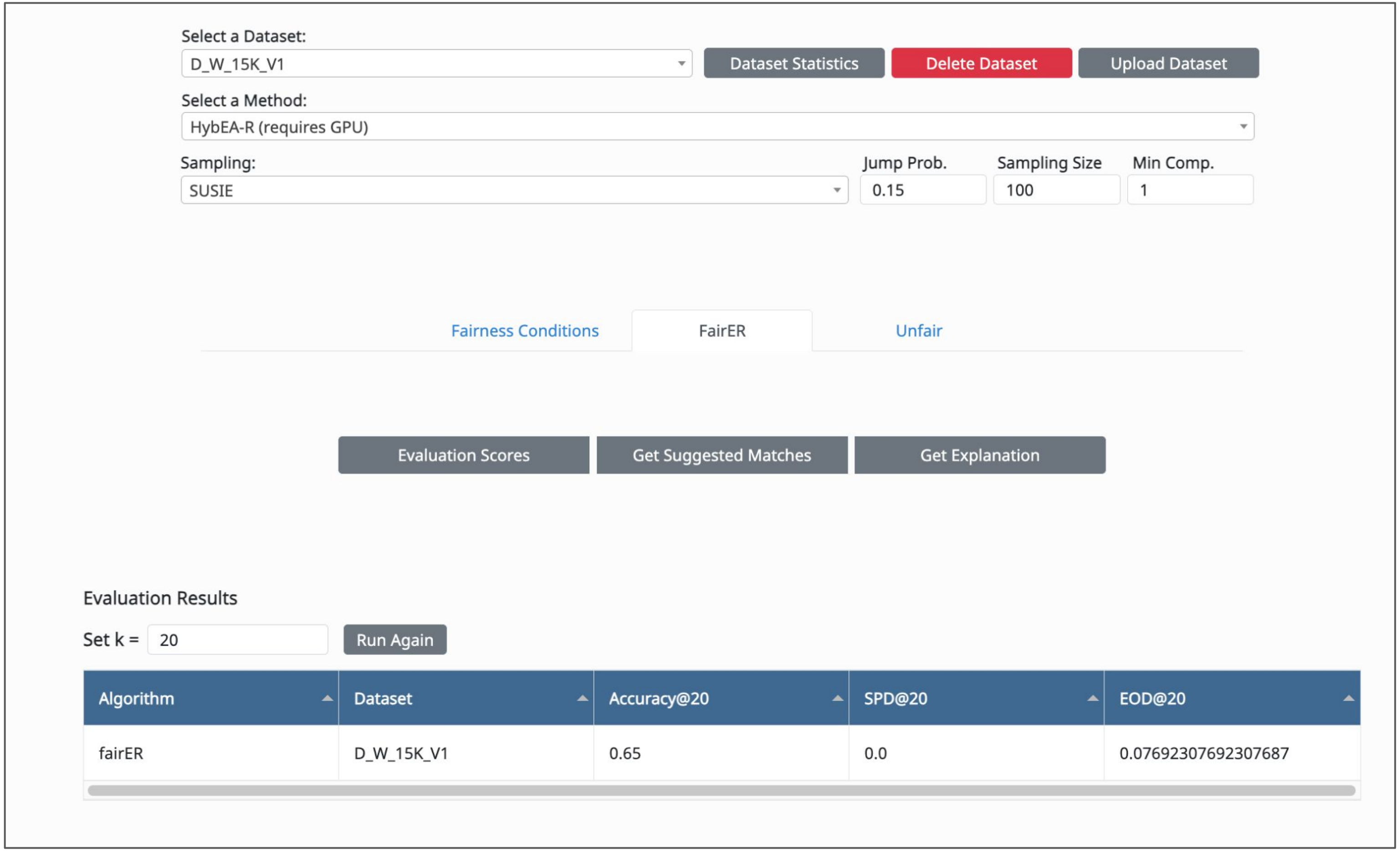
4 KG and
6 Tabular
Datasets

7 KG
methods

1 Tabular
Method



Settings and Evaluation Scores



Visual Explanations on Suggested Matches

Protected Candidates

KG_1 id	KG_2 id
Andrei_Tar...	Q853
Perryton,_...	Q975601
Myazedi	Q6946825
Carl_Fried...	Q84790
~~Herbert_~~...	~~Q84790~~
The_Loves...	Q3988126
Myitkyina_...	Q11917400
Herbert_...	Q76641
~~Carl_Fried~~...	~~Q76641~~
Nonprofit_...	Q163740

First Prev Next Last

Non Protected Candidates

KG_1 id	KG_2 id
Missionary	Q219477
Oh!_Sabella	Q75569
~~Oh!_Sabella~~	~~Q1093636~~
Do_Aur_D...	Q4746547
Dino_Risi	Q53034
Strange_R...	Q7621451
~~Oh!_Sabella~~	~~Q3827879~~
Castle	Q23413
Taste_of_C...	Q55210
Joseph_Lu...	Q553882

First Prev Next Last

Suggested Matches

Protected Non protected

Rank	KG_1 id	KG_2 id	Matching Score
#0	Andrei_Tarkovs...	Q853	0.89963221549...
#1	Missionary	Q219477	0.89219719171...
#2	Perryton,_Texas	Q975601	0.82994306087...
#3	Oh!_Sabella	Q75569	0.85428726673...
#4	Myazedi	Q6946825	0.75230240821...
#5	Do_Aur_Do_Pa...	Q4746547	0.78974771499...
#6	Carl_Friedrich_...	Q84790	0.70037758350...
#7	Dino_Risi	Q53034	0.78842961788...
#8	The_Loves_of_...	Q3988126	0.67013180255...
#9	Strange_Relati...	Q7621451	0.78800439834...

First Prev 1 2 3 4 5 Next Last

Conclusions

- We experimentally **assessed** the **different degrees** of **structural and factual heterogeneity** exhibited by real KGs
- HybEA is able to **adapt** to the different degrees of factual and structural heterogeneity exhibited by real KGs
    - Outperforms **11 SOTA EA methods** achieving a **16% average relative** improvement of Hits@1, **ranging from 3.6% up to 40%** in all **7 monolingual datasets**, with some datasets that can now be considered as ***solved***, while also in **3 multilingual datasets** with a **statistically significance difference**
- We introduced an **exploration-based sampling method**, that allows sampling KGs of adjustable levels of **structural diversity** of KGs
    - Demonstrate that SOTA KGE-based EA methods exhibit indirect bias against smaller, less connected regions of benchmark datasets.
    - HybEA-R is the most robust method to structural bias

Future Work

- **Explain HybEA matching decisions** to enhance transparency and trustworthiness
    - **coarse-grained:** it makes it easy to see the provenance of the returned matches (i.e., matches returned by the structural vs the factual component)
    - **fine-grained:** the factual model could report the most significant attributes of aligned entities, while the structural model their most significant relations

- **Extend SUSIE** to consider not only structural diversity, but also factual diversity (i.e., in literal values)

- In the context of **temporal evolving KGs** and **streaming settings**
    - **Our semi-supervised framework** is a first step towards this direction by proposing reliable **pseudo-labels**
    - **Extend HybEA** also with **incremental node embedding modules**

Publications

1. Nikolaos Fanourakis, Vasilis Efthymiou, Dimitris Kotzinos, Vassilis Christophides: **Knowledge graph embedding methods for entity alignment: experimental review.** Data Min. Knowl. Discov 2023
2. Nikolaos Fanourakis, Vasilis Efthymiou, Vassilis Christophides, Dimitris Kotzinos, Evaggelia Pitoura, Kostas Stefanidis: **Structural Bias in Knowledge Graphs for the Entity Alignment Task**. ESWC 2023
3. Nikolaos Fanourakis, Christos Kontousias, Vasilis Efthymiou, Vassilis Christophides, Dimitris Plexousakis: **Fairness-Aware and Explainable Entity Resolution**. ISWC (Posters/Demos/Industry) 2023
4. Nikolaos Fanourakis, Fatia Lekbour, Guillaume Renton, Vasilis Efthymiou, Vassilis Christophides: **HybEA: Hybrid Models for Entity Alignment**. CoRR abs/2407.02862 2024 (under review)

THANK YOU!