EVALUATION INOVATIONS IN KNOWLEDGE GRAPH COMPLETION

2025-09-04

Presenter : Sooho Moon

MINDS

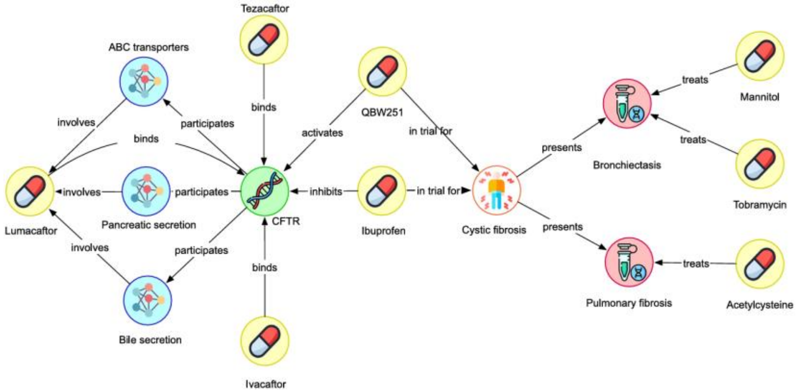
INDEX

- Knowledge Graph Completion
- Evaluation Procedure
- Filtered Setting
- Breaking Ties
- The Open World

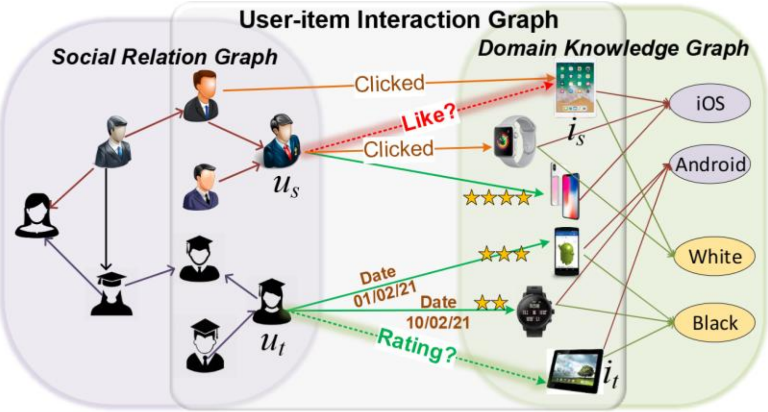
KNOWLEDGE GRAPH COMPLETION

■ What is a Knowledge Graph(KG)?

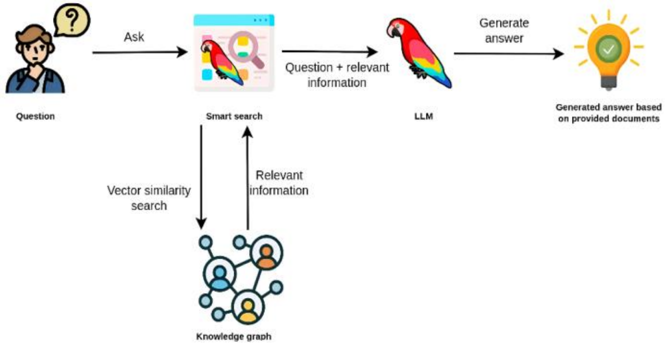
- ☐ Heterogeneous graph consisting of **entities**(nodes) and **relations**(edges)
- ☐ **Triple** = (head entity, relation, tail entity)
- ☐ Rich representation → applied in various domains



<Drug Discovery>



<Recommender System>

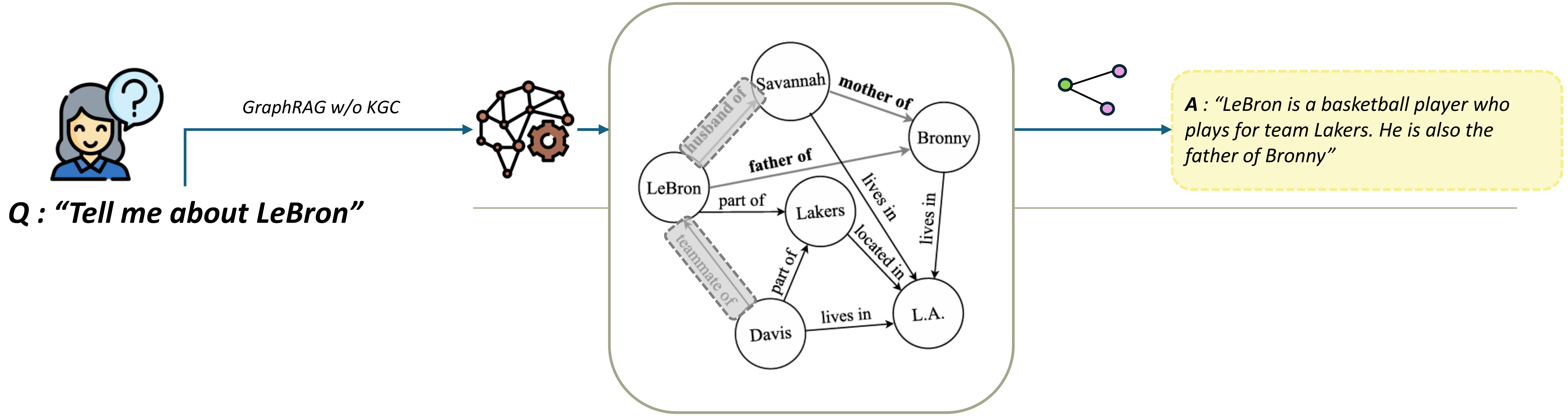


<GraphRAG>

KNOWLEDGE GRAPH COMPLETION

■ Missing Facts in KGs

- ☐ In Freebase and DBpedia, more than 66% of the person entities are missing a birthplace
- ☐ Greatly hinder the performance of systems that rely upon it

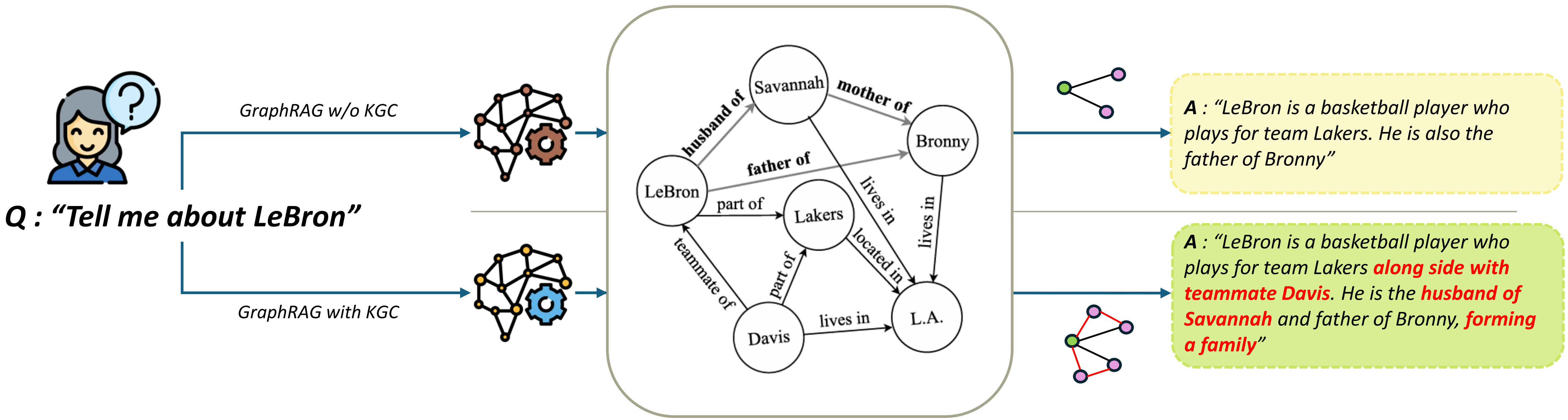


KNOWLEDGE GRAPH COMPLETION

■ Knowledge Graph Completion(KGC)

- ☐ KGC model to automatically fill missing links has been extensively studied
- ☐ Normally two tasks exist, link prediction(entity prediction) and relation prediction

link prediction : (h, r, ?) or (?, r, t) → predict "?"



EVALUATION PROCEDURE

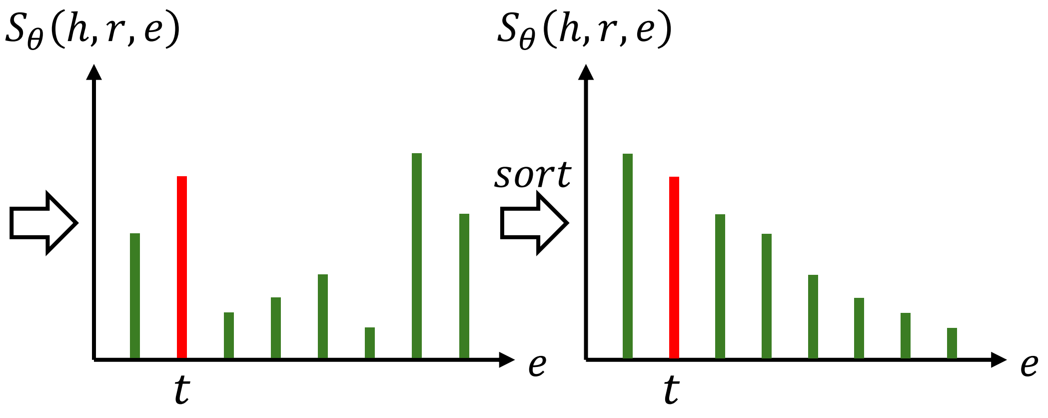
■ Evaluation Procedure of KGC Models

$(\mathrm{h}, \mathrm{r}, \mathrm{t}) \in G^{tst}$

$S_\theta : (\mathrm{h}, \mathrm{r}, \mathrm{t}) \rightarrow score \in \mathbb{R}$

query : $(\mathrm{h}, \mathrm{r}, ?)$

$S_\theta(h, r, e)\ for\ all\ e\ in\ \mathcal{E}$



$$MR = \frac{1}{Q} \sum_i rank_i$$

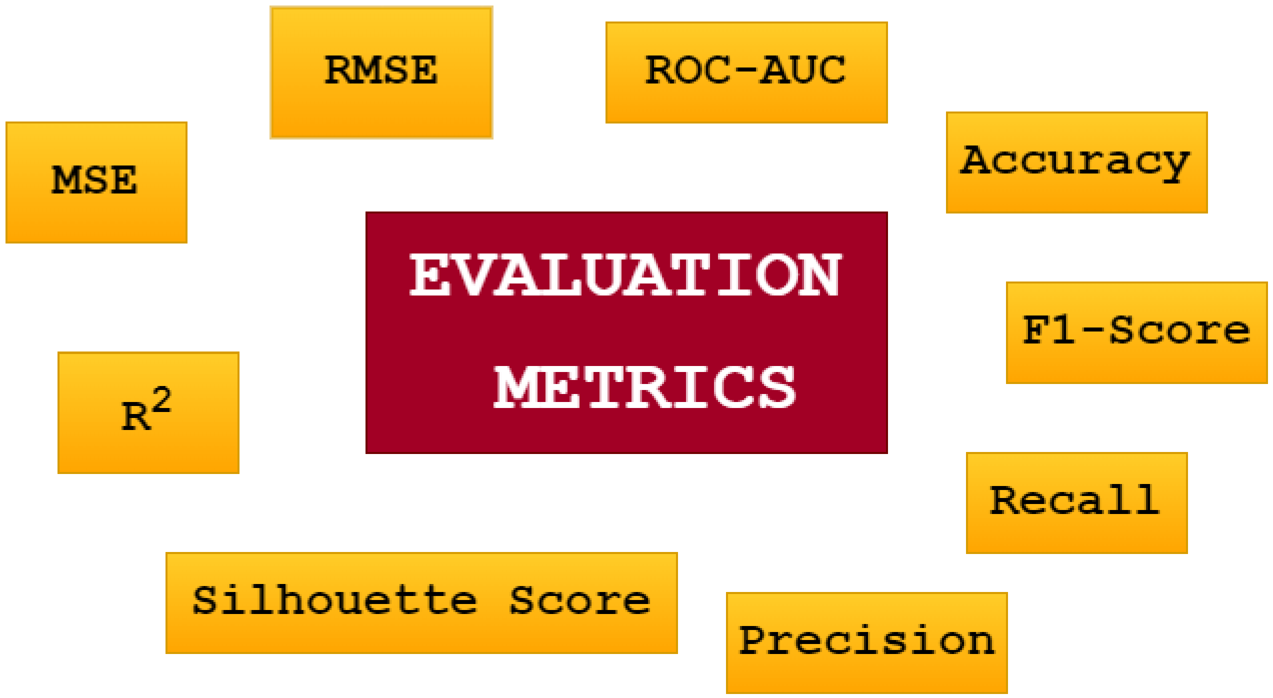
$$MRR = \frac{1}{Q} \sum_i 1/rank_i$$

$$Hits@k = \frac{1}{Q} \sum_i \mathbb{I}[rank_i \leq k]$$

EVALUATION PROCEDURE

■ The Judge : Metric

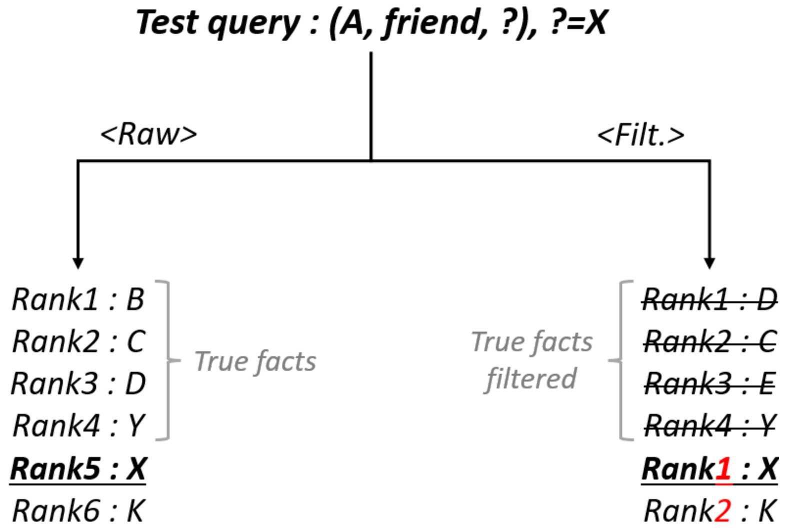
- ☐ *'Which is the best model?'*
- ☐ Communities use conventional metrics according to the task environment



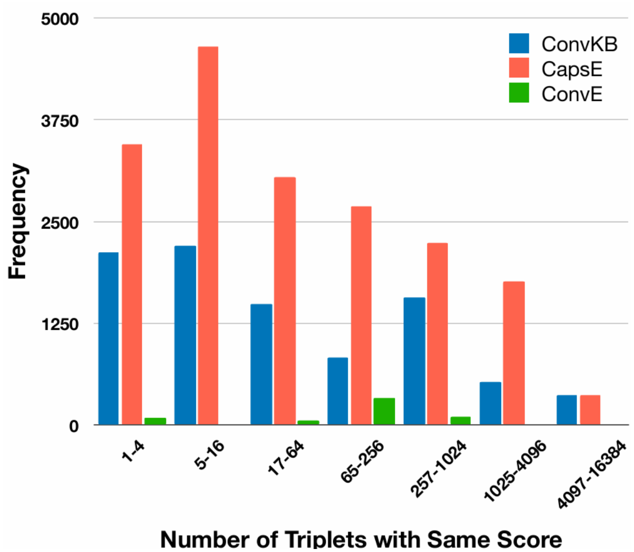
EVALUATION PROCEDURE

■ Flaws of Metrics & Protocols

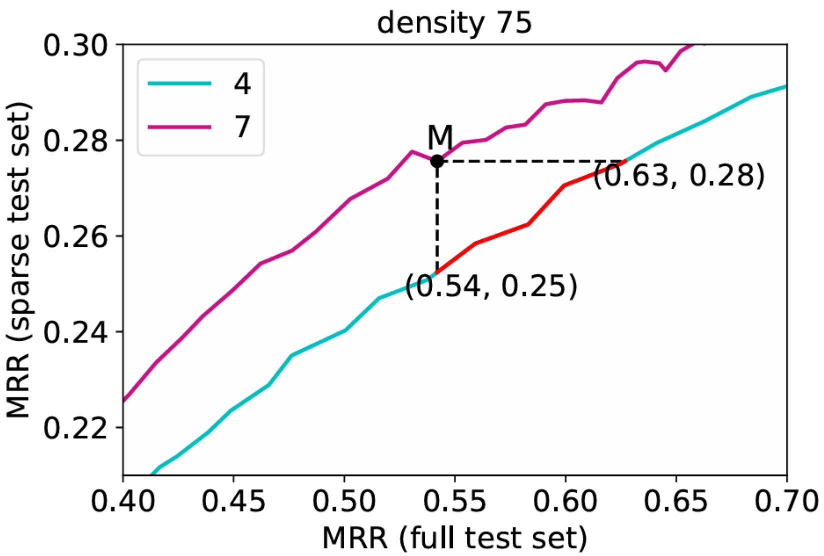
Filtered setting'13



Breaking ties'20



Open world problem'22



FILTERED SETTING

- Translating Embeddings
- Filtered Setting
- Experiments
- Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

Translating Embeddings for Modeling Multi-relational Data

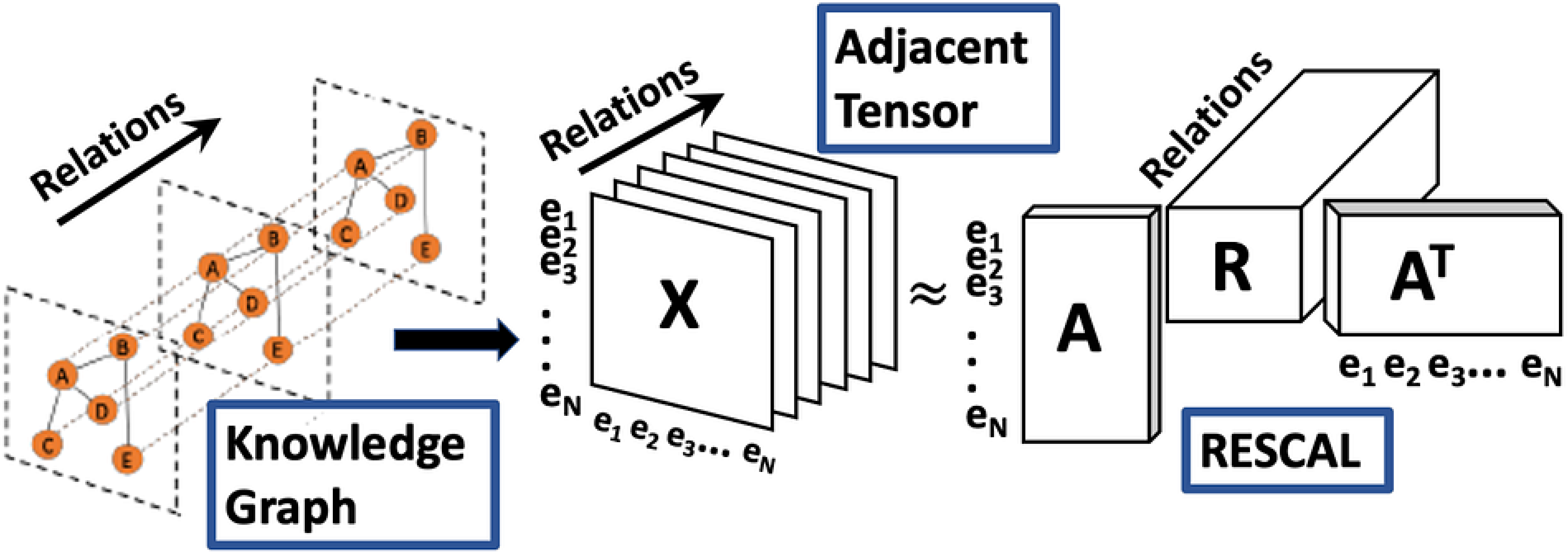
Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán
Université de Technologie de Compiègne – CNRS
Heudiasyc UMR 7253
Compiègne, France
{bordesan, nusunier, agarciad}@utc.fr

Jason Weston, Oksana Yakhnenko
Google
111 8th avenue
New York, NY, USA
{jweston, oksana}@google.com

TRANSLATING EMBEDDINGS

■ Tensor Factorization for KGC

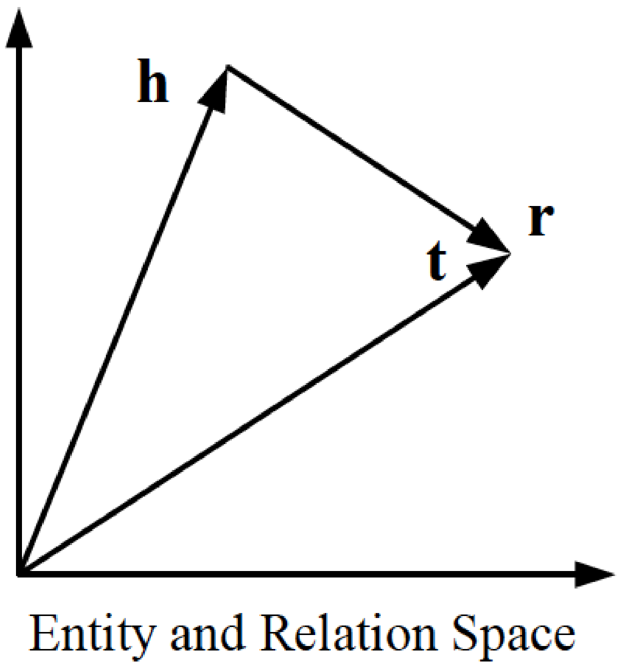
- ☐ Popular approach for modeling multi-relational data
- ☐ Expensive model training (e.g., RESCAL)

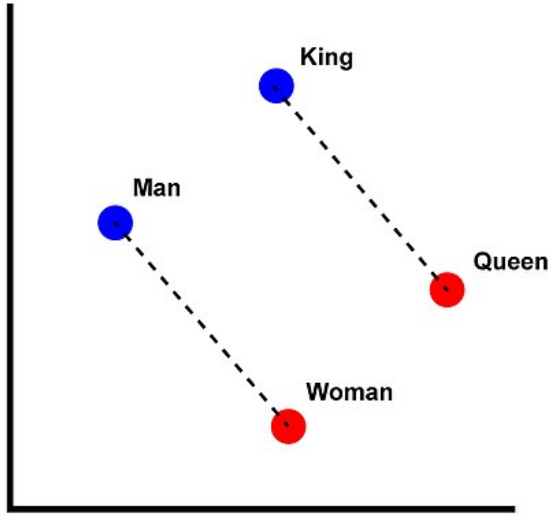


TRANSLATING EMBEDDINGS

■ Simple Vector Addition

- ☐ View entity, relation as individual vector
- ☐ Optimize $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$ for every triple in KG : "**Trans**lating **E**mbeddings"



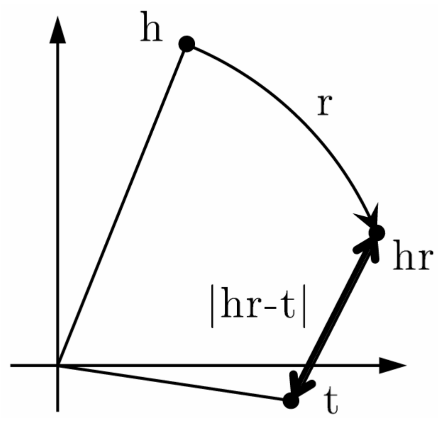


Look similar?

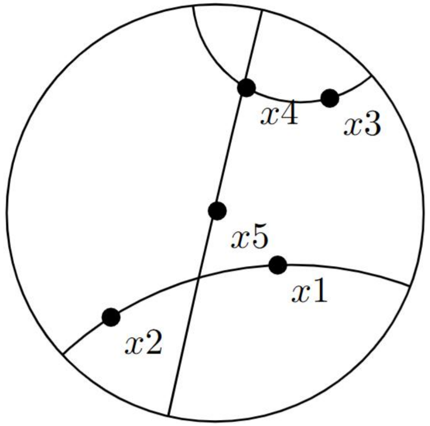
TRANSLATING EMBEDDINGS

■ Revolutionized KGC

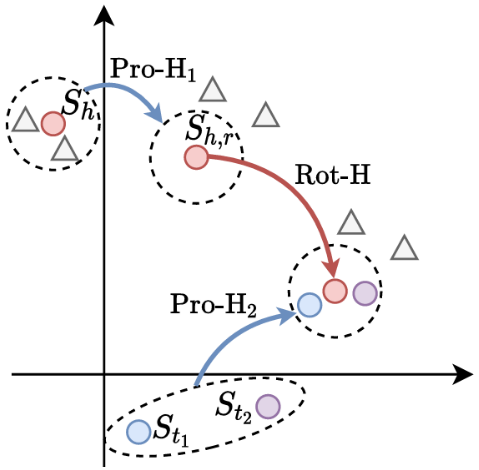
- ☐ Significant drop in training cost
- ☐ Simple design & mathematical motivation → numerous extensions



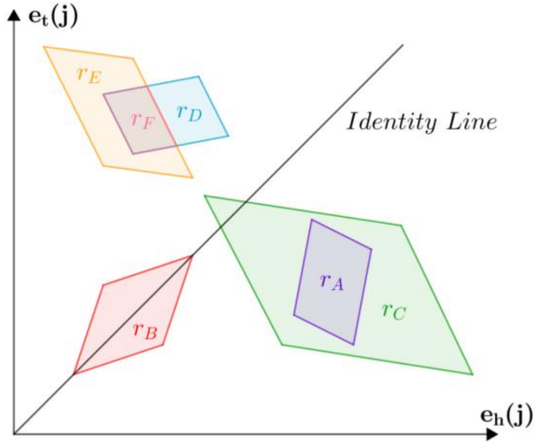
RotatE (2019)



MuRP (2019)



HousE (2022)

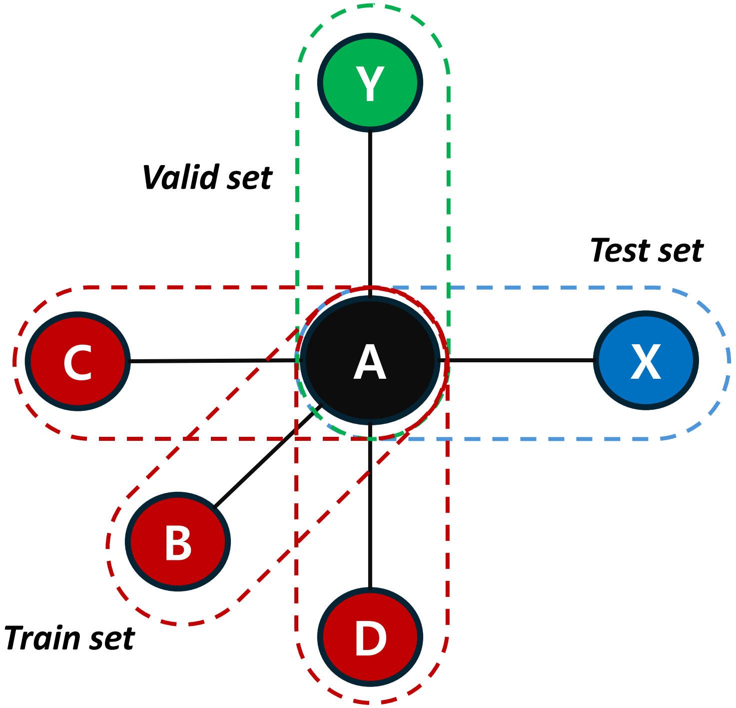


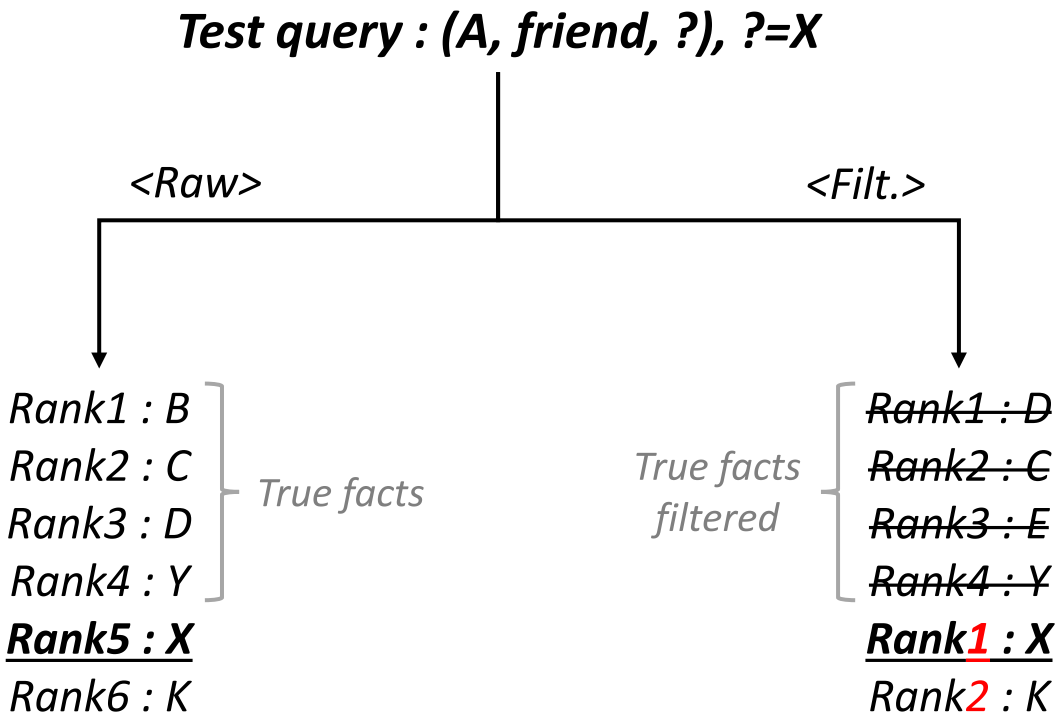
ExpressivE (2023)

FILTERED SETTING

■ Introduction to Filtered Setting

☐ Raw : with out filtered setting, Filt. : with filtered setting





FILTERED SETTING

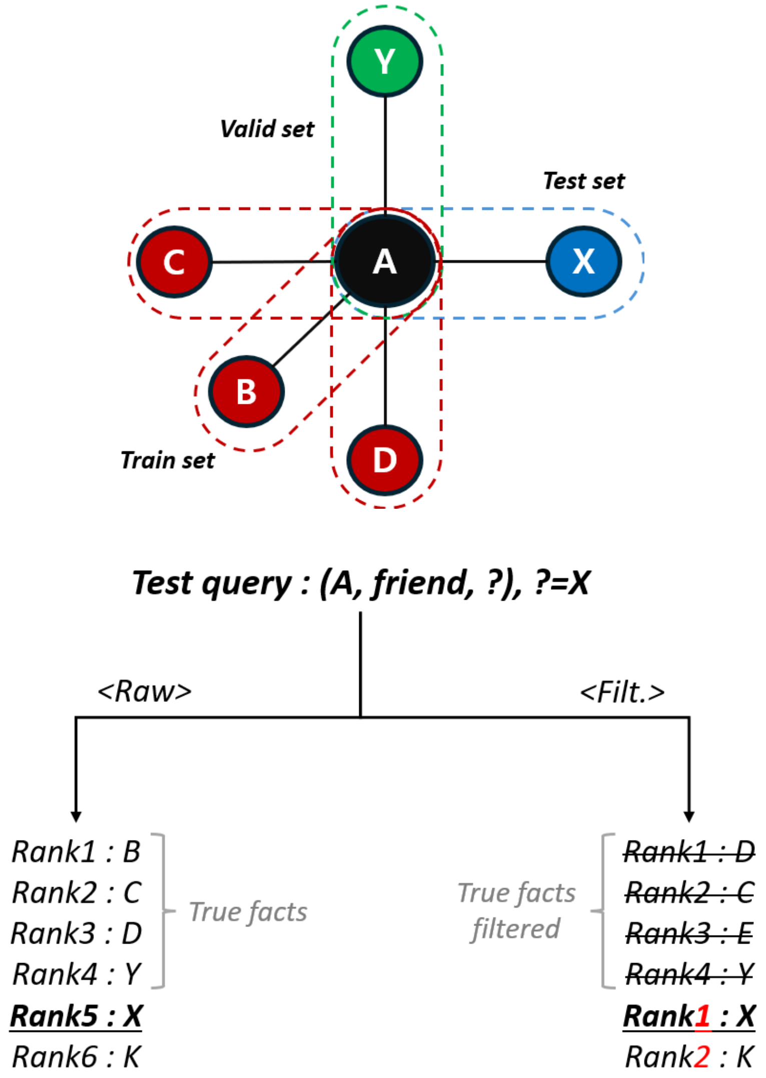
Experiment Results

DATASET	WN				FB15K				FB1M	
METRIC	MEAN RANK		HITS@10 (%)		MEAN RANK		HITS@10 (%)		MEAN RANK	HITS@10 (%)
Eval. setting	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Filt.*	*Raw*	*Raw*
Unstructured [2]	315	304	35.3	38.2	1,074	979	4.5	6.3	15,139	2.9
RESCAL [11]	1,180	1,163	37.2	52.8	828	683	28.4	44.1	-	-
SE [3]	1,011	985	68.5	80.5	273	162	28.8	39.8	22,044	17.5
SME(LINEAR) [2]	545	533	65.1	74.1	274	154	30.7	40.8	-	-
SME(BILINEAR) [2]	526	509	54.7	61.3	284	158	31.3	41.3	-	-
LFM [6]	469	456	71.4	81.6	283	164	26.0	33.1	-	-
TransE	**263**	**251**	**75.4**	**89.2**	**243**	**125**	**34.9**	**47.1**	**14,615**	**34.0**

CONCLUSION

■ Simple yet Obvious

- ☐ Presence of external answers should not occupy rank position
- ☐ Simple yet obvious idea



■ Influence on the Community

- ☐ Became the ‘default’ protocol
- ☐ Main reason : ‘no reason not to’

BREAKING TIES

- Background
- Problem Search
- New Protocol
- Experiments
- Extended Discussion
- Conclusion

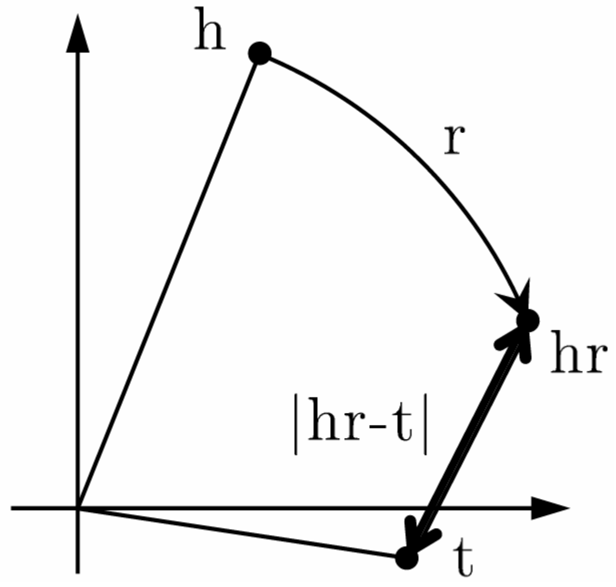
A Re-evaluation of Knowledge Graph Completion Methods

Zhiqing Sun[1*] **Shikhar Vashishth**[1,2*] **Soumya Sanyal**[2*]
Partha Talukdar[2] **Yiming Yang**[1]
[1] Carnegie Mellon University, [2] Indian Institute of Science
{zhiqings,svashish,yiming}@cs.cmu.edu
{soumyasanyal,ppt}@iisc.ac.in

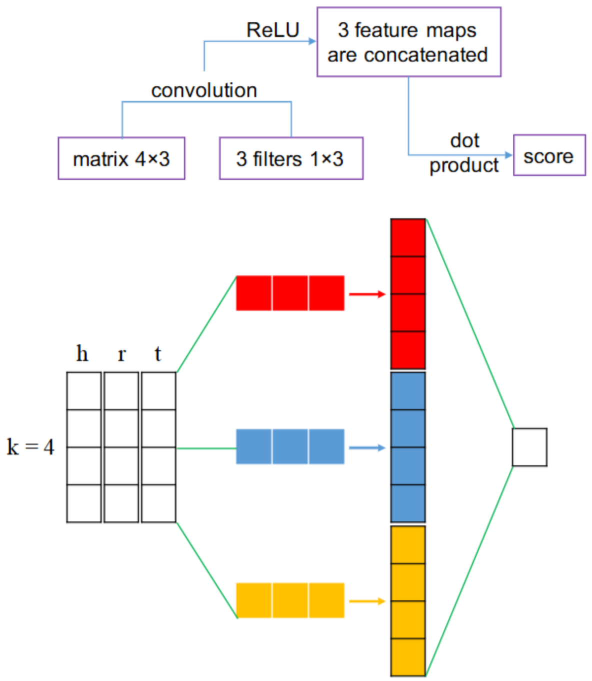
BACKGROUND

■ Many Types of KGC Models

- ☐ SOTA KGC methods have been published in top conferences in recent years



Embedding based



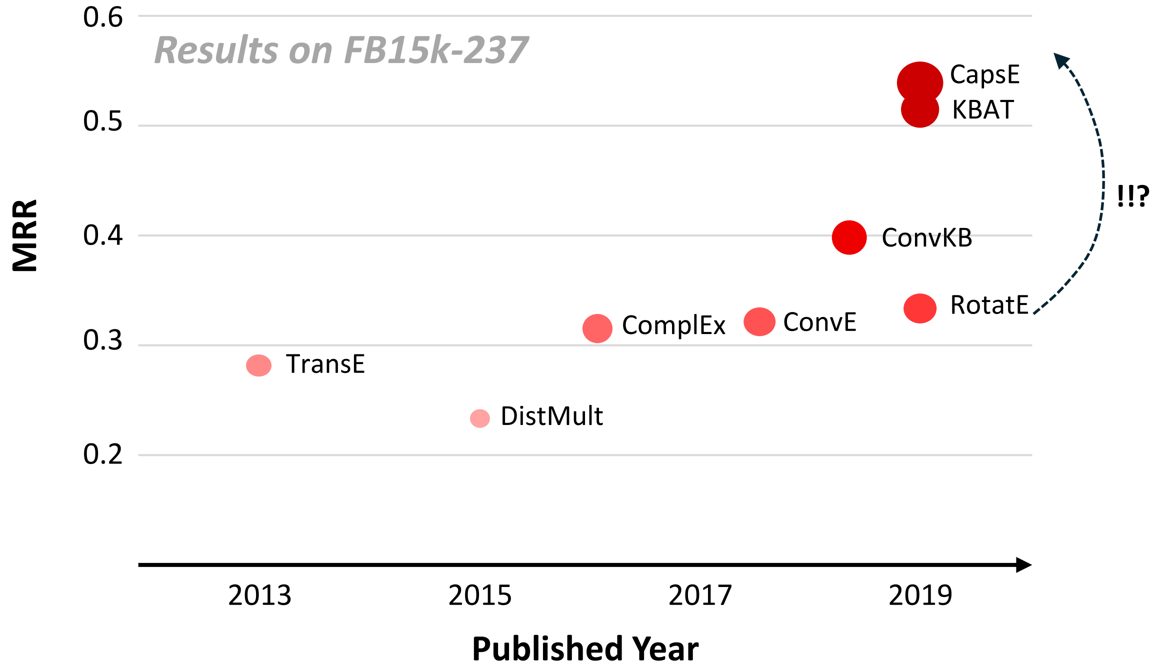
Neural Network based

BACKGROUND

■ Casting Doubt on Reported Results

- ☐ Several **NN based models** reported odd performance on FB15k-237 dataset
- ☐ Outlier performance **not consistent** across different datasets

① Odd gains

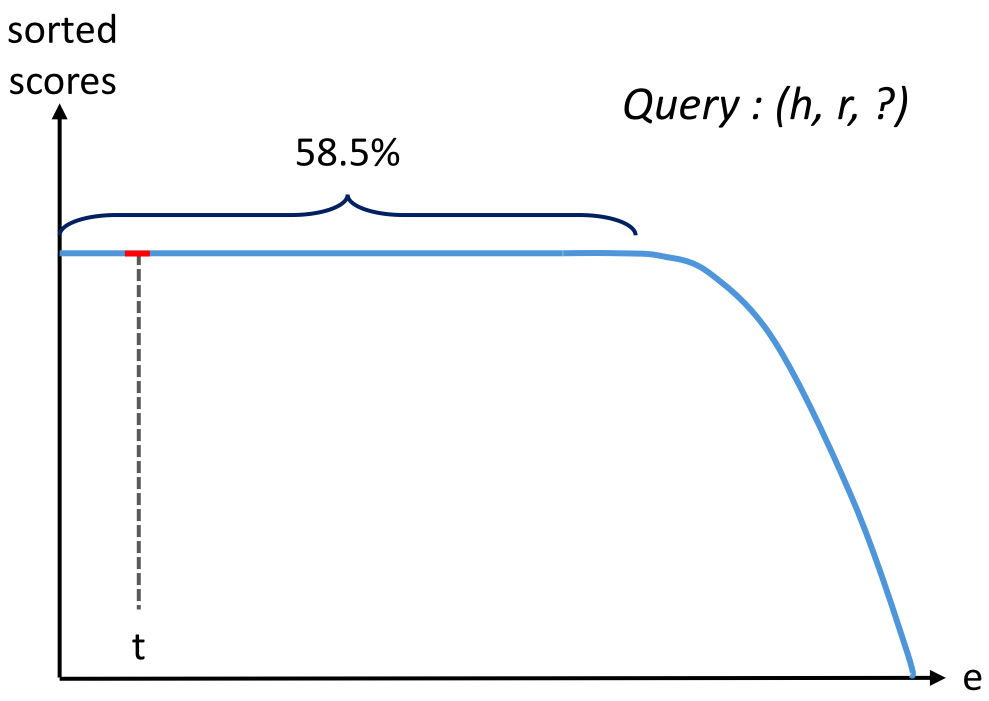


② Inconsistency

	FB15k-237	WN18RR
ConvE	.325	.430
RotatE	.338 (+4.0%)	.476 (+10.6%)
TuckER	.358 (+10.2%)	.470 (+9.3%)
ConvKB	.396 (+21.8%)	.248 (-42.3%)
CapsE	.523 (+60.9%)	.415 (-3.4%)
KBAT	.518 (+59.4%)	.440 (+2.3%)
TransGate	.404 (+24.3%)	.409 (-4.9%)

PROBLEM SEARCH

■ (Obs.1) Too Many Same Scores



$$r(t) = count[score(e) > score(t)] + 1$$

No discriminative power

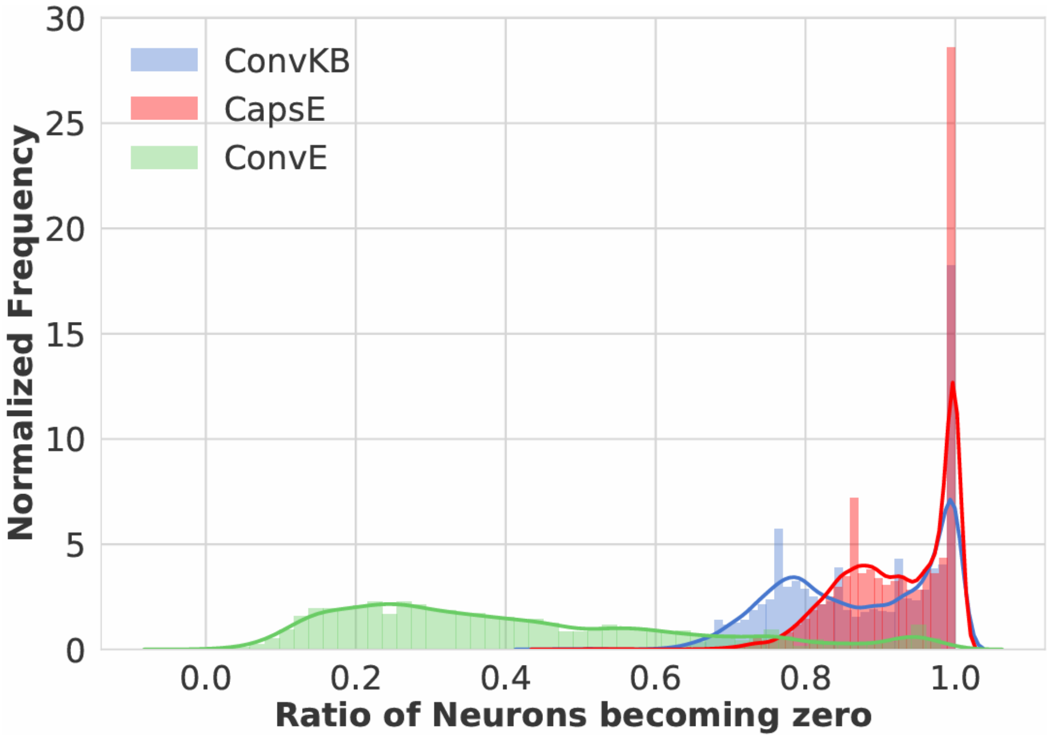
$$r(t) = position_{score}[t]$$

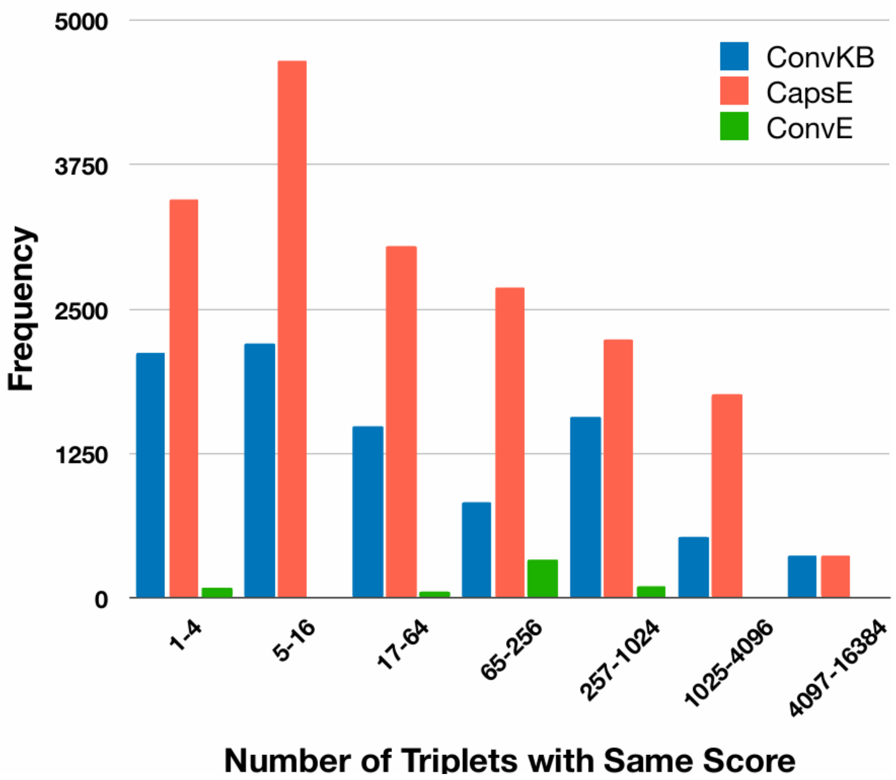
Dependent on the sorting algorithm

PROBLEM SEARCH

■ (Obs.2) Dead Neurons After Activation

☐ Reason for so many same scores ($\sigma = ReLU$)

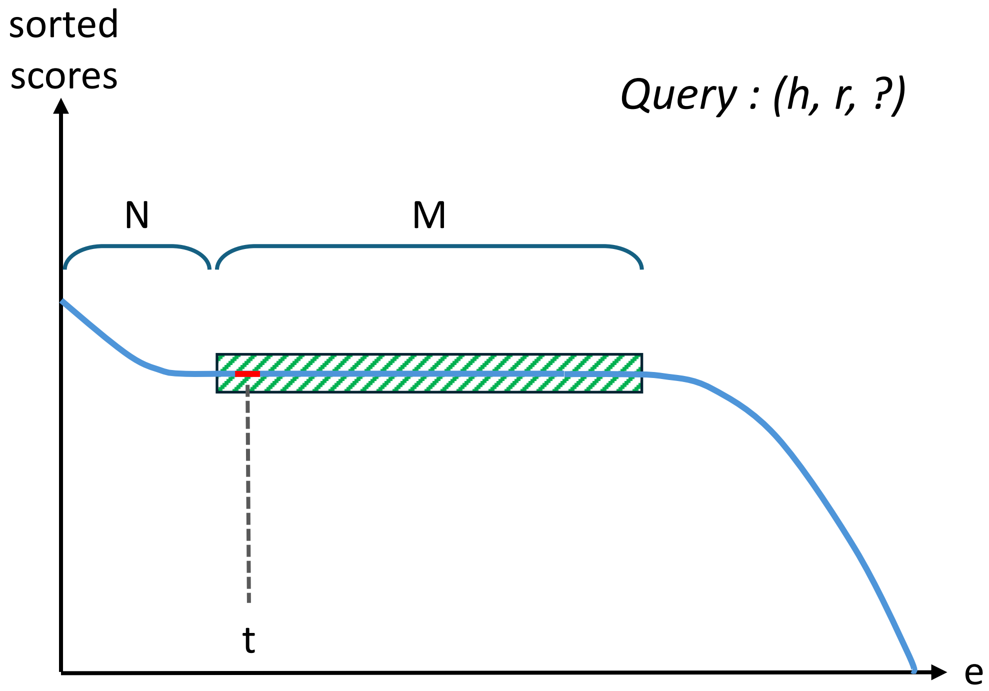




NEW PROTOCOL

■ Random Protocol for Tied Scores

- ☐ Models the realistic situation when ties occurred during knowledge extraction



TOP $r(t) = \mathrm{N} + 1$

$\boxed{\textbf{RANDOM} \quad r(t) = random[N+1, N+M]}$

BOTTOM $r(t) = \mathrm{N} + \mathrm{M}$

EXPERIMENTS

■ FB15k237 : Random VS Top VS Bottom

	Reported			RANDOM			TOP			BOTTOM		
	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑
ConvE	.325	244	.501	.324 ± .0	285 ± 0	.501 ± .0	.324	285	.501	.324	285	.501
RotatE	.338	177	.533	.336 ± .0	178 ± 0	.530 ± .0	.336	178	.530	.336	178	.530
TuckER	.358	-	.544	.353 ± .0	162 ± 0	.536 ± .0	.353	162	.536	.353	162	.536
ConvKB	.396	257	.517	.243 ± .0	309 ± 2	.421 ± .0	.407 (+.164)	246 (-63)	.527 (+.106)	.130 (-.113)	373 (+64)	.383 (-.038)
CapsE	.523	303	.593	.150 ± .0	403 ± 2	.356 ± .0	.511 (+.361)	305 (-99)	.586 (+.229)	.134 (-.016)	502 (+99)	.297 (-.059)
KBAT	.518†	210†	.626†	.157 ± .0	270 ± 0	.331 ± .0	.157	270	.331	.157	270	.331

*KBAT reimplemented due to test data leakage

EXPERIMENTS

■ WN18RR : Random VS Top VS Bottom

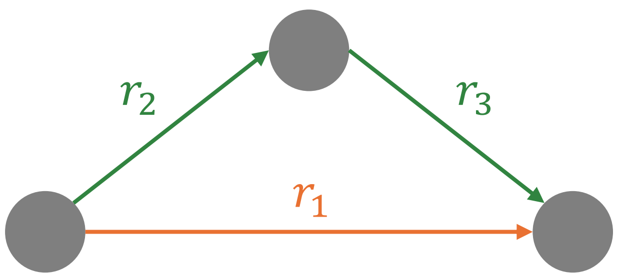
	Reported			RANDOM			TOP			BOTTOM		
	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑
ConvE	.43	4187	.52	.444 ± .0	4950 ± 0	.503 ± .0	.444	4950	.503	.444	4950	.503
RotatE	.476	3340	.571	.473 ± .0	3343 ± 0	.571 ± .0	.473	3343	.571	.473	3343	.571
TuckER	.470	-	.526	.461 ± .0	6324 ± 0	.516 ± .0	.461	6324	.516	.461	6324	.516
ConvKB	.248	2554	.525	.249 ± .0	3433 ± 42	.524 ± .0	.251 (+.002)	1696 (-1737)	.529 (+.005)	.164 (-.085)	5168 (+1735)	.516 (-.008)
CapsE‡	.415	719	.560	.415 ± .0	718 ± 0	.559 ± .0	.415	718	.559	.323 (-.092)	719 (+1)	.555 (-.004)
KBAT	.440†	1940†	.581†	.412 ± .0	1921 ± 0	.554 ± .0	.412	1921	.554	.412	1921	.554

*KBAT reimplemented due to test data leakage

EXTENDED DISCUSSION

■ Potential Problem : Rule-Based Model

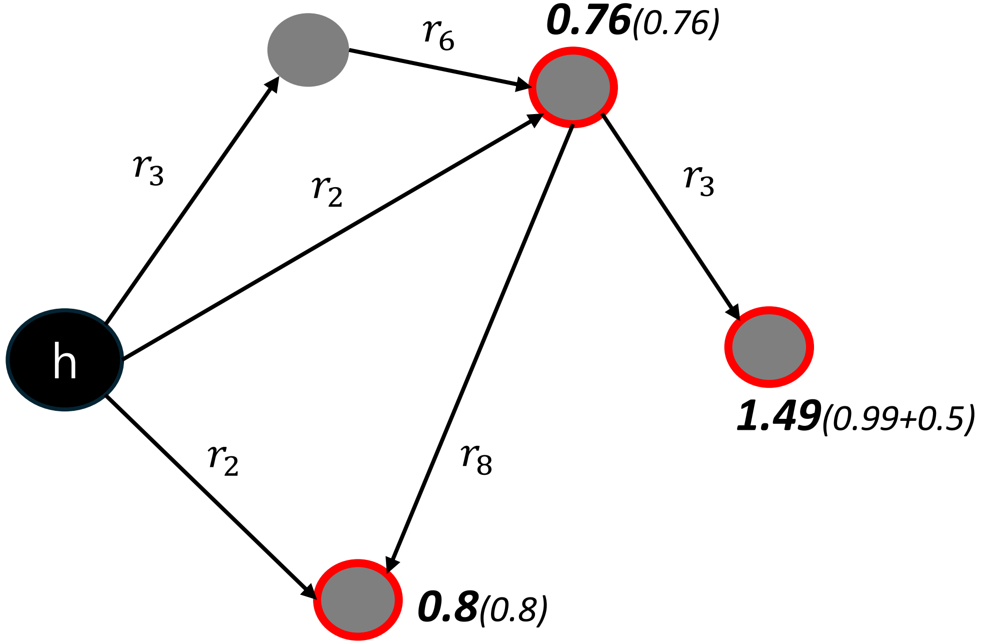
$head$
$: [body]\ (confidence)$



Mined rules

$$r_1 : r_2 \rightarrow r_3\ (0.99)$$
$$r_1 : r_2 \rightarrow r_8\ (0.8)$$
$$r_1 : r_3 \rightarrow r_6\ (0.76)$$
$$r_1 : r_3 \rightarrow r_6 \rightarrow r_3 (0.5)$$
$$r_1 : r_4 \rightarrow r_3\ (0.1)$$
$$r_2 : r_1 \rightarrow r_2 \rightarrow r_3 (0.8)$$
$$r_2 : r_1 \rightarrow r_3\ (0.6)$$
$$\vdots$$

$Query : (h, r_1, ?)$



EXTENDED DISCUSSION

Potential Tied Scores

$$\boldsymbol{score\ of\ (h, r, e)} \xrightarrow{\emptyset \neq r\{h \mapsto e\}} \sum_{path \in r\{h \mapsto e\}} confidence(path)$$

$$\boldsymbol{score\ of\ (h, r, e)} \xrightarrow[\emptyset = r\{h \mapsto e\}]{} 0$$

Discrete scoring mechanism

Rule-based

$$\boldsymbol{score\ of\ (h, r, e)} \xrightarrow{\textit{Distance based}} |\boldsymbol{h} \times \boldsymbol{r} - \boldsymbol{t}|$$

$$\boldsymbol{score\ of\ (h, r, e)} \xrightarrow[\textit{Score based}]{} \boldsymbol{h} \times \boldsymbol{r} \times \boldsymbol{t}$$

Continuous scoring mechanism

Emb-based

EXTENDED DISCUSSION

■ Gap Between Reported Results of Rule-Based Models

Method	Family				Kinship				UMLS				WN18RR				FB15k-237				YAGO3-10			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
TransE†	0.450	22.1	-	87.4	0.310	0.9	-	84.1	0.690	52.3	-	89.7	0.230	2.2	-	52.4	0.294	18.9	-	46.5	0.360	25.1	-	58.0
DistMult†	0.540	36.0	-	88.5	0.354	18.9	40.0	75.5	0.391	25.6	44.5	66.9	0.430	39.0	44.0	49.0	0.241	15.5	26.3	41.9	0.340	24.3	-	53.3
ComplEx†	0.810	72.7	-	94.6	0.418	24.2	49.9	81.2	0.411	27.3	46.8	70.0	0.440	41.0	46.0	51.0	0.247	15.8	27.5	42.8	0.340	24.8	-	54.9
ConvE†	-	-	-	-	-	-	-	-	-	-	-	-	0.430	40.0	44.0	52.0	0.320	21.6	-	50.1	0.440	35.5	-	61.6
R-GCN†	-	-	-	-	-	-	-	-	-	-	-	-	0.402	34.5	43.7	49.4	0.273	18.2	30.3	45.6	-	-	-	-
TuckER†	-	-	-	-	0.630	46.2	69.8	86.3	0.732	62.5	81.2	90.9	0.470	44.3	48.2	52.6	**0.358**	**26.6**	**39.4**	**54.4**	-	-	-	-
RotatE†	0.860	78.7	-	93.3	0.651	50.4	75.5	93.2	0.744	63.6	82.2	**93.9**	0.476	42.8	49.2	**57.1**	0.338	24.1	37.5	53.3	0.490	40.2	-	**67.0**
AMIE+†	0.541	48.7	59.2	59.6	0.416	23.7	49.3	83.0	0.417	25.3	51.9	68.5	0.192	19.0	19.5	19.6	0.118	8.8	13.1	17.3	0.259	23.1	28.4	30.3
MLN†	-	-	-	-	0.351	18.9	40.8	70.7	0.688	58.7	75.5	86.9	-	-	-	-	-	-	-	-	-	-	-	-
PathRank†	-	-	-	-	0.369	27.2	41.6	67.3	0.197	14.8	21.4	25.2	0.189	17.1	20.0	22.5	0.087	7.4	9.2	11.2	-	-	-	-
NeuralLP†	0.880	80.1	-	98.5	0.302	16.7	33.9	59.6	0.483	33.2	56.3	77.5	0.381	36.8	38.6	40.8	0.237	17.3	25.9	36.1	-	-	-	-
DRUM†	0.890	82.6	-	**99.2**	0.334	18.3	37.8	67.5	0.548	35.8	69.9	81.4	0.382	36.9	38.8	41.0	0.238	17.4	26.1	36.4	-	-	-	-
NLIL†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.250	-	-	32.4	-	-	-	-
CTP†	-	-	-	-	0.335	17.7	37.6	70.3	0.404	28.8	43.0	67.4	-	-	-	-	-	-	-	-	-	-	-	-
RLogic†	0.880	81.3	-	97.2	0.580	43.4	-	87.2	0.710	56.6	-	93.2	0.470	44.3	-	53.7	0.310	20.3	-	50.1	0.360	25.2	-	50.4
NCRL*	0.910	85.2	-	99.3	0.640	49.0	-	92.9	0.790	65.9	-	95.1	0.670	56.3	-	85.0	0.300	20.9	-	47.3	0.380	27.4	-	53.6
NCRL†	0.806	74.0	86.0	90.1	0.537	38.6	62.6	83.4	0.562	45.9	61.6	71.5	0.263	23.2	28.0	31.3	0.141	7.5	17.0	26.6	0.012	0.3	1.3	2.8
RNNLogic†	0.860	79.2	-	95.7	0.639	49.5	73.1	92.4	0.745	63.0	83.3	92.4	0.455	41.4	47.5	53.1	0.288	20.8	31.5	44.5	0.379	30.2	42.1	53.3
TCLM†	**0.985**	**98.1**	**98.9**	99.1	**0.686**	**54.3**	**79.5**	**95.3**	**0.808**	**73.0**	**87.1**	92.8	**0.483**	**44.7**	**49.7**	55.2	0.311	23.0	34.0	47.2	**0.505**	**43.4**	**54.7**	63.5
TCLM (w/o IC)†	0.984	97.9	**98.9**	99.0	0.684	54.1	79.1	95.1	0.788	70.0	86.2	92.8	0.462	42.7	47.6	53.6	0.309	22.8	33.8	46.7	0.430	34.3	47.4	59.1

CONCLUSION

■ Subtle yet Crucial

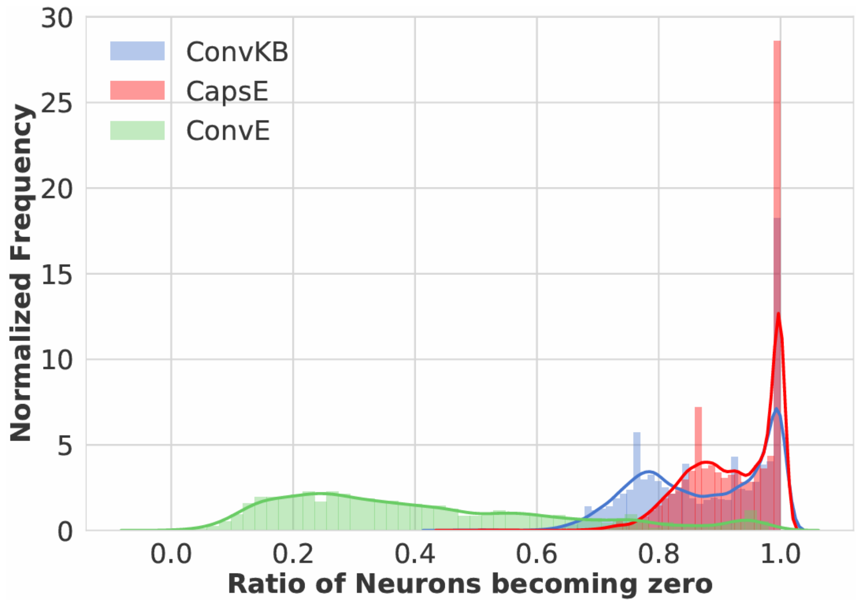
- ☐ Observed anomalous metric results

■ Finding the Cause & Remedy

- ☐ Rank ties resulting in metric inflation
- ☐ Random protocol for reasonable evaluation

■ Influence on the Community

- ☐ Quite recognized among researchers
- ☐ Not fully adopted compared to the filtered setting



	Reported			RANDOM		
	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑
ConvE	.325	244	.501	.324 ± .0	285 ± 0	.501 ± .0
RotatE	.338	177	.533	.336 ± .0	178 ± 0	.530 ± .0
TuckER	.358	-	.544	.353 ± .0	162 ± 0	.536 ± .0
ConvKB	.396	257	.517	.243 ± .0	309 ± 2	.421 ± .0
CapsE	.523	303	.593	.150 ± .0	403 ± 2	.356 ± .0
KBAT	.518†	210†	.626†	.157 ± .0	270 ± 0	.331 ± .0

THE OPEN WORLD

- Background
- Problem Definition
- Theoretical Approach
- Experiments
- Discussion

NEURAL INFORMATION PROCESSING SYSTEMS

Rethinking Knowledge Graph Evaluation Under the Open-World Assumption

Haotong Yang[1,2] **Zhouchen Lin**[1,2,3]* **Muhan Zhang**[2,4]*
[1]Key Lab of Machine Perception (MoE),
School of Intelligence Science and Technology, Peking University
[2]Institute for Artificial Intelligence, Peking University
[3]Peng Cheng Laboratory
[4]Beijing Institute for General Artificial Intelligence
{haotongyang, zlin, muhan}@pku.edu.cn

BACKGROUND

■ Inherent and Philosophical Question

Flight search system

Is there a **direct flight from USA to North Korea**?

Searching DB…
Info not found

"No, direct flight doesn't exist"

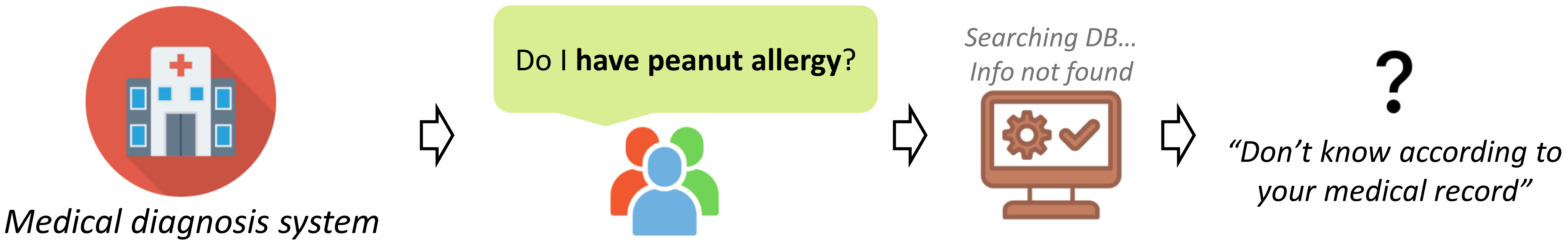
fact not present → fact false

system's knowledge is <u>complete</u> = the world is <u>closed</u>

Closed World Assumption

BACKGROUND

■ Inherent and Philosophical Question



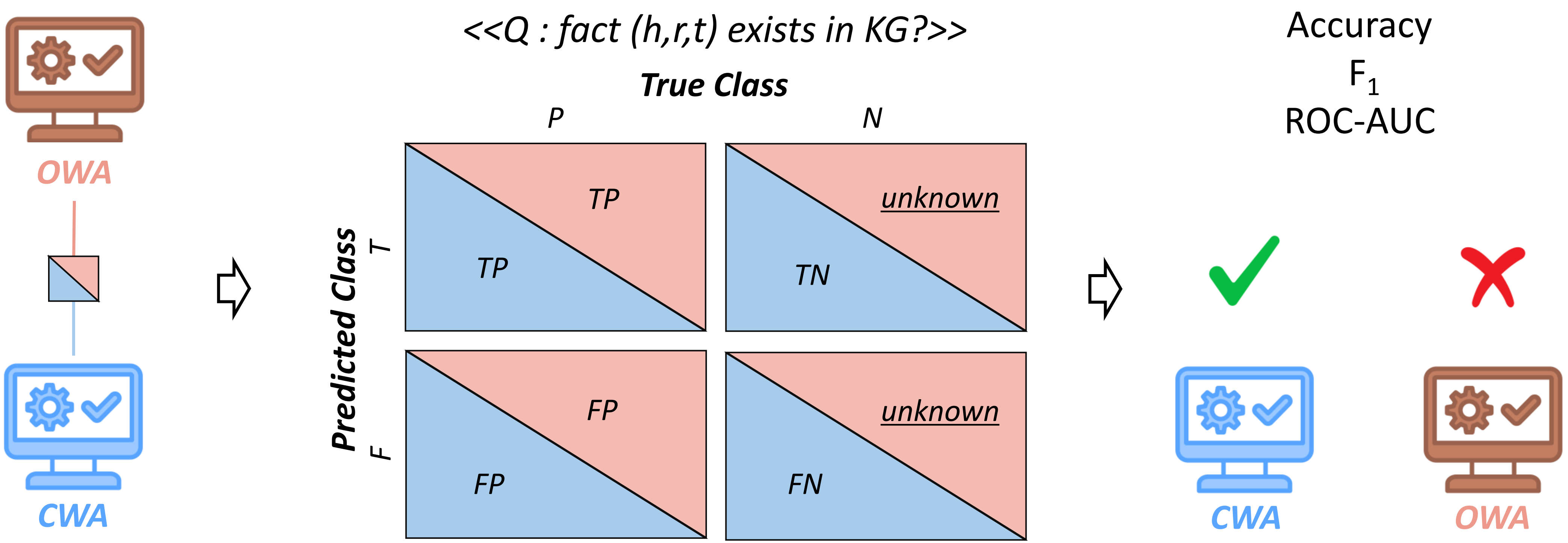
fact not present → fact unknown

system's knowledge is <u>incomplete</u> = the world is <u>open</u>

Open World Assumption

BACKGROUND

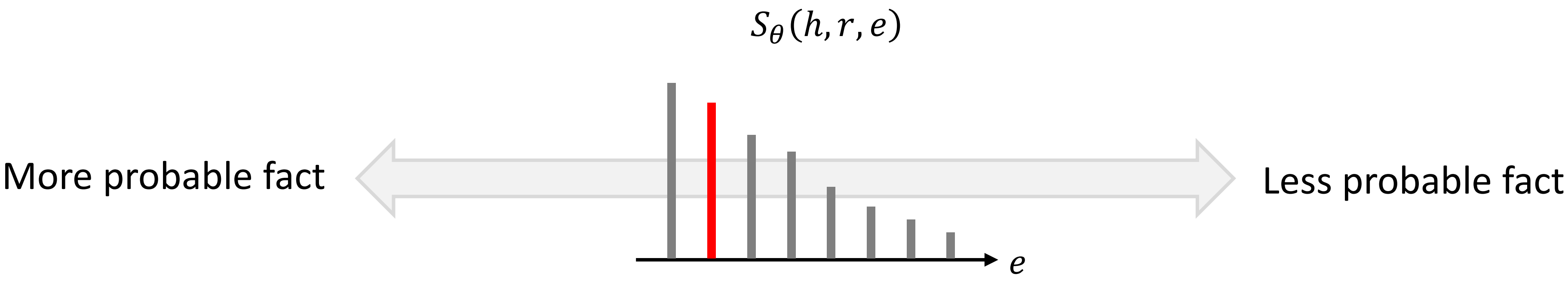
Crucial Definition Before Evaluation



BACKGROUND

■ Rank-Based Metrics

- ☐ FN/TN not required
- ☐ Only interested in putting true positives in front of unknowns



■ KGC and Rank-Based Metrics

- ☐ Finding new knowledge = KGC's main purpose → **KGC follows OWA**
- ☐ Exclusive use of rank-based metrics (MR, MRR, Hits@k)

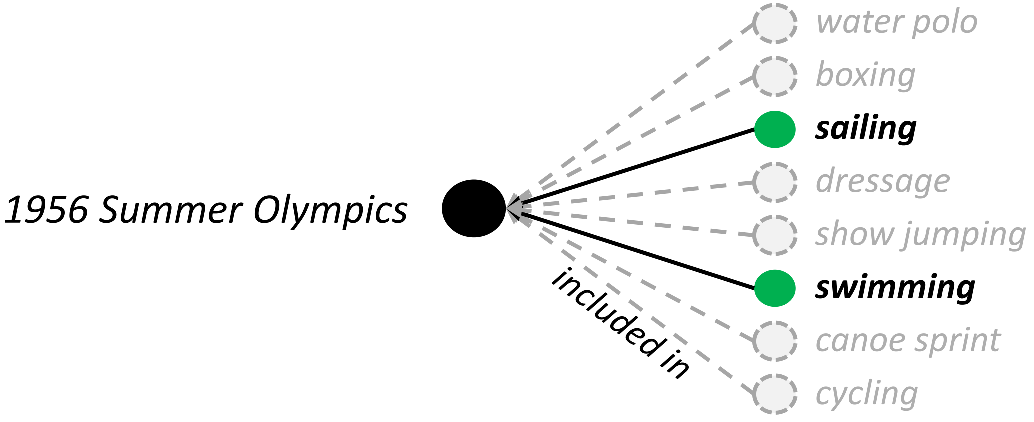
PROBLEM DEFINITION

Problem with Missing Answers

- (?, included in, 1965 Summer Olympics); **existing answers** = swimming, sailing

BetaE
[Ren and Leskovec' 20]

MRR (w/o c) = ***0.2***
MRR (with c) = 1.0



	Test Answers		Missing Answers					
	swimming	sailing	**water polo**	**boxing**	**dressage**	**show jumping**	**canoe sprint**	**cycling**
ranking w/o c	5	5	1	2	3	4	7	9
ranking with c	1	1	1	1	1	1	3	4

w/o c : with out correction, with c : with correction

PROBLEM DEFINITION

Alleviating the Sparsity

- Augmentation via Gen-AI : feasible? **No**

Resource issue

FB15k-237 fact check

1500 years

$200 million

Practicality issue

(Obama, born in, Hawaii)

T / F

(dantrolene, synergistic interaction with, diltiazem)

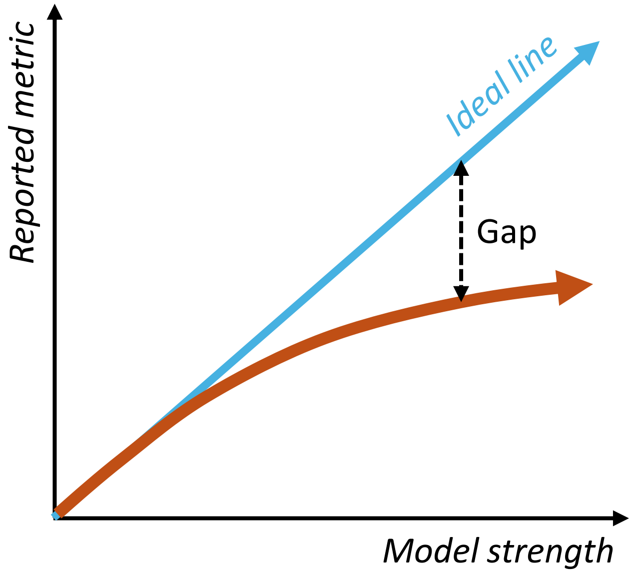
T? / F?

	FB15k-237	WN18RR	YAGO3-10	DDB14	Hetionet
$\|E\|$	14541	40943	123182	9203	45158
$\|R\|$	237	11	37	14	23
$\|E\|^2\|R\|$	**50B**	**18B**	**561B**	**1B**	**47B**

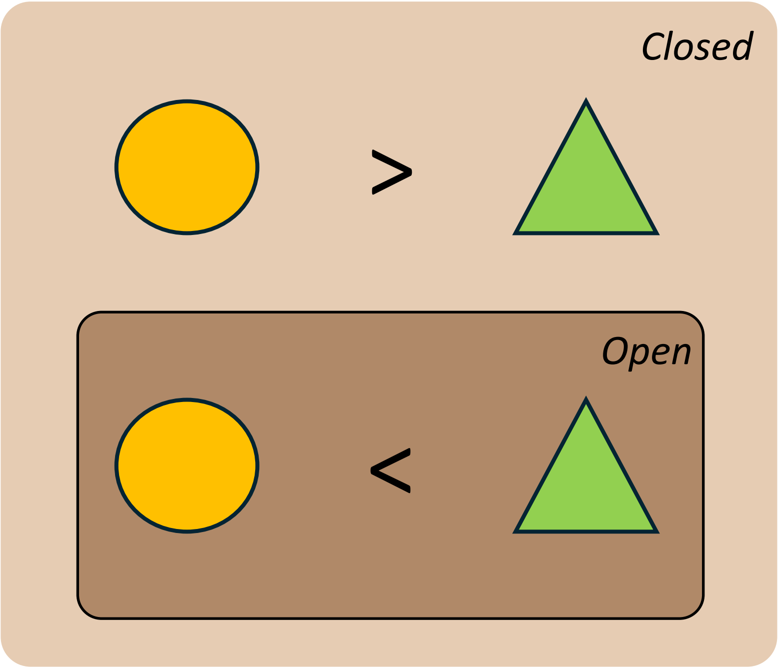
PROBLEM DEFINITION

■ Two Problems Affecting the KGC Evaluation

1) Metric degradation



2) Metric inconsistency



PROBLEM DEFINITION

■ The Ultimate Question

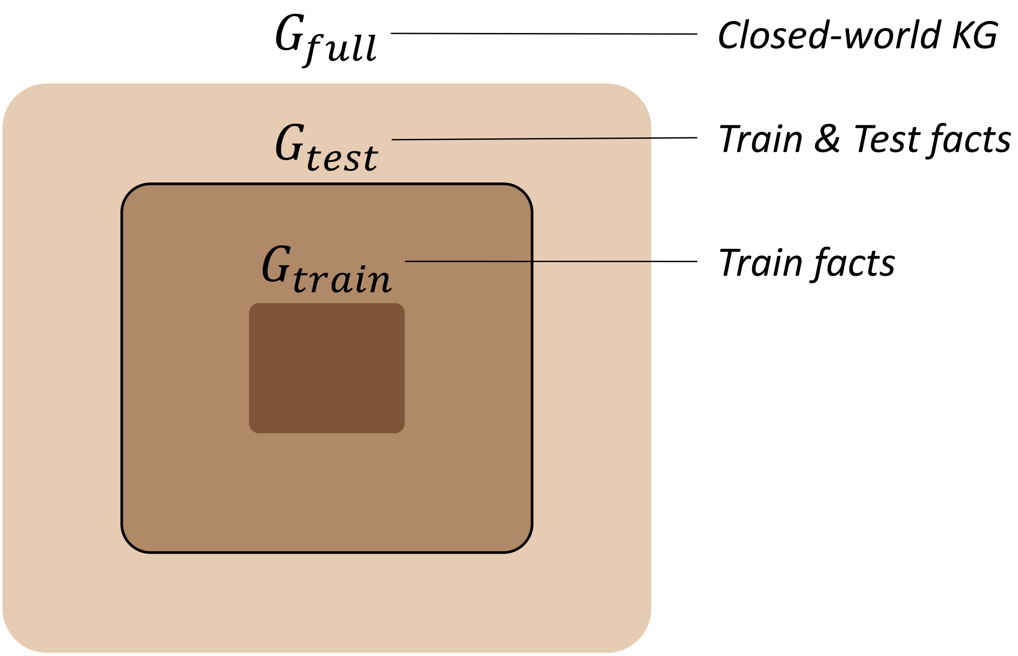
Full test facts

$$G_{full} \backslash G_{train}$$

Sparse test facts

$$G_{test} \backslash G_{train}$$

"Are evaluation from sparse test facts and full test facts consistent?"



THEORETICAL APPROACH

■ Two Source of Randomness

1) Sparsity of KG : β

$$X : r(e_h, e_t) \in G_{full} \backslash G_{train} \\ \wedge\ r(e_h, e_t) \notin G_{test} \backslash G_{train}$$

$$P(X) = \beta = 1 - \alpha$$

2) Model strength : ℓ

$$Y : KGC \cong classification\ model$$

$$P(Y) = \ell$$

independence

$$\mathfrak{M} \sim P(G_{test} \backslash G_{train}, X, Y)$$

THEORETICAL APPROACH

■ Expectation Degradation

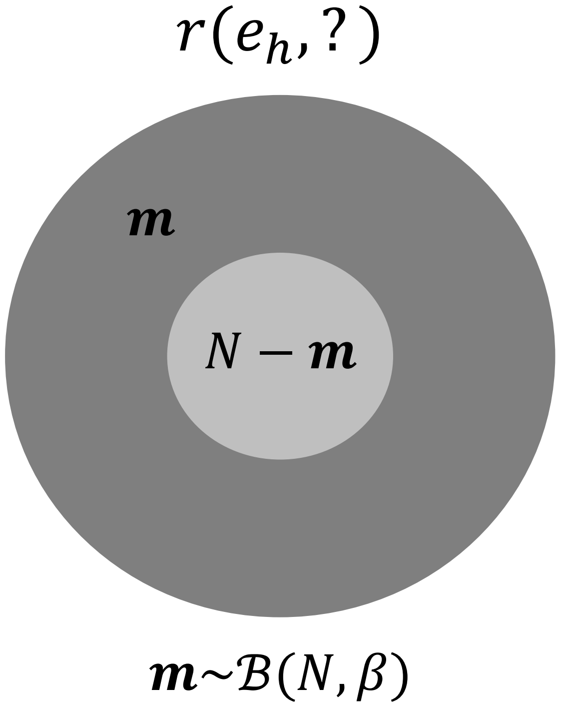
MRR : $f(r) = r$

Hits@k : $f(r) = 1 \; if \; r \leq k \; else \; \infty$

$$\mathfrak{M} = \frac{1}{N - \boldsymbol{m}} \sum_{i=1}^{N-\boldsymbol{m}} \frac{1}{f(\boldsymbol{r}(e_i))}$$

$$\Downarrow$$

$$\hat{\mathbb{E}} = \mathbb{E}(\mathfrak{M}) = \frac{1}{\beta(N+1)} \sum_{k=0}^{N} \frac{1}{f(k+1)} \left(1 - \hat{\Phi}(k)\right)$$

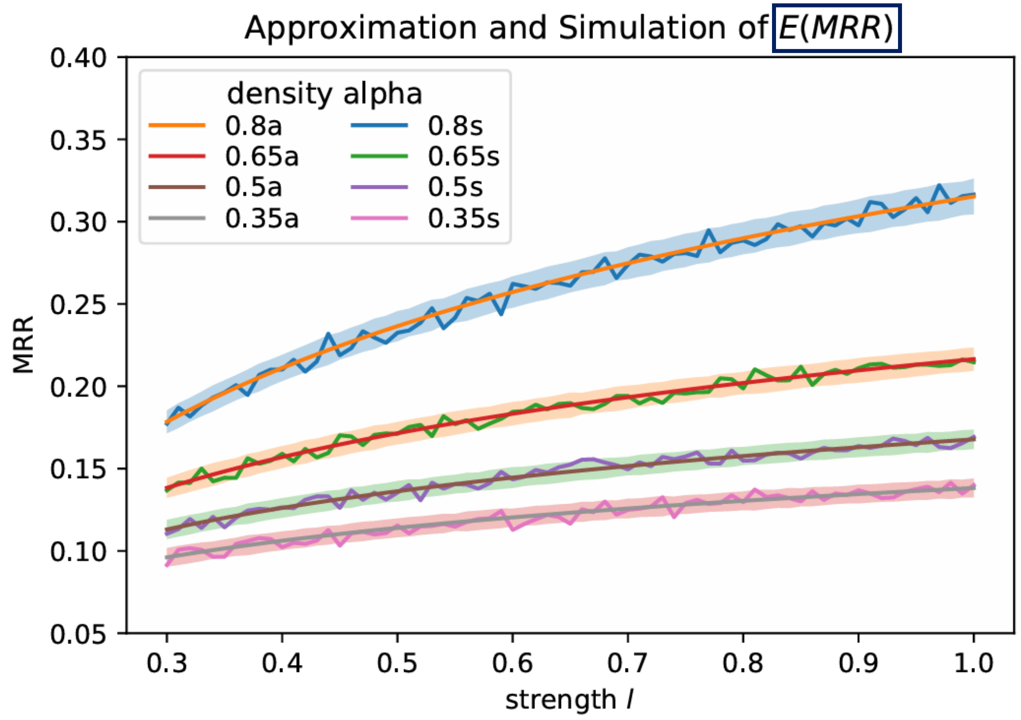


THEORETICAL APPROACH

Expectation Degradation

$$\boxed{\hat{\mathbb{E}}(\mathrm{MRR})} \approx \frac{\ln(\ell) + \ln(\beta) + \ln(N+2) + \gamma}{\beta(N+1)} := \tilde{\mathbb{E}}$$

- MRR will behave like a **log function w.r.t the growth of ℓ**
- **Perfect model can't achieve MRR=1** due to sparsity β



THEORETICAL APPROACH

■ Inconsistency Due to High Variance

$\mathcal{M}_1 = \ell$ < $\mathcal{M}_2 = \ell + \Delta\ell$

$$P[\mathfrak{M}(\mathcal{M}_1) \geq \mathfrak{M}(\mathcal{M}_2)] \xrightarrow{\text{reject}} N_q \geq \mathcal{O}((1/\Delta\ell)^2)$$

# of test queries quadratically grows w.r.t $1/\Delta\ell$

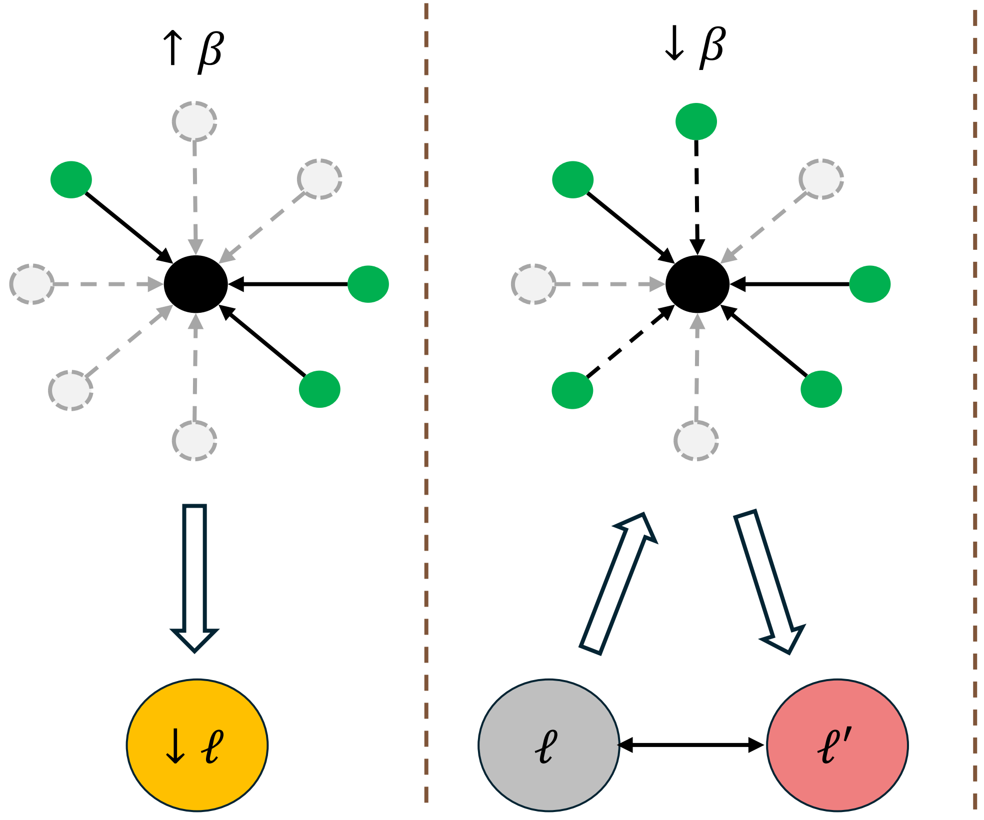
$\Downarrow$

Close model strengths demand cautious comparison

THEORETICAL APPROACH

■ Considering Correlation Between β and ℓ

① Model under-trained ② Previously complemented

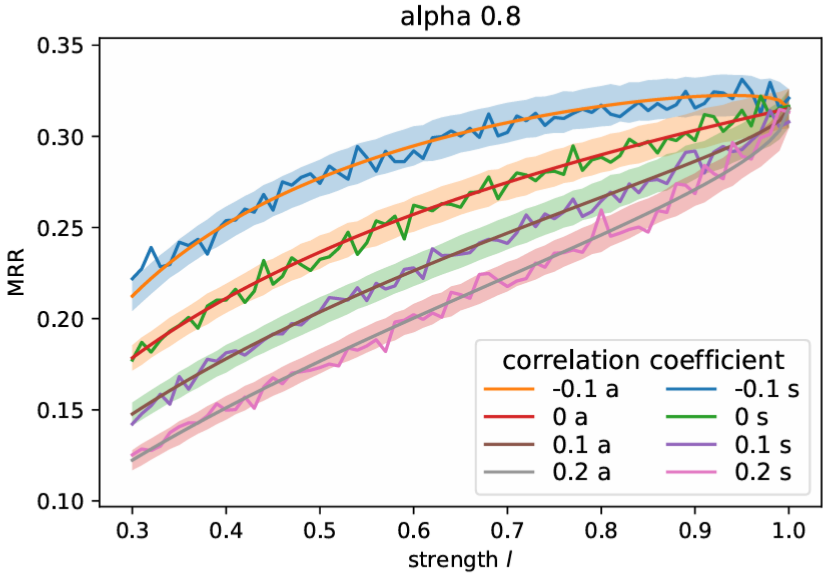


Negative correlation between β and ℓ

According to Corollary 4.5 in the paper,

MRR favor models with **smaller ρ instead of larger ℓ**

More severe inconsistency!!!



THEORETICAL APPROACH

Need for New Metric

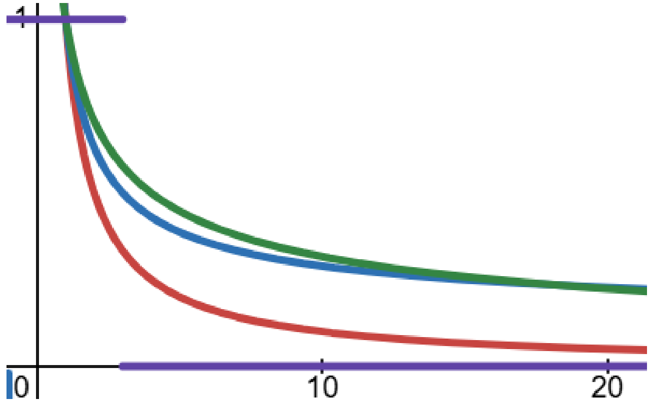
- The only changeable element : transform function f, **but how?**

Focus-on-top

- MRR : $f(r) = r$
- Hits@k : $f(r) = \mathbb{I}[r \le k]$

- log-MRR : $f(r) = \log_2(r+1)$
- p-MRR : $f(r) = r^p (0 < p < 1)$

Less focus-on-top



Let $g(\boldsymbol{r}) = \boldsymbol{r}/f(\boldsymbol{r}), \boldsymbol{r} \in \mathbb{N}_+$

$$\frac{\mathrm{d}\hat{\mathbb{E}}(\mathfrak{M})}{\mathrm{d}\ell} = \frac{1}{l\beta(N+1)} \mathbb{E}_{k \sim \mathcal{B}(N+1, \ell\beta)}\, g(k)$$

↓ focus-on-top → ↓ degradation & ↓ inconsistency

More credibility to the conclusion

EXPERIMENTS

■ Dataset, Models, Metrics Setup

Dataset

Family-tree

# entities	6004
# relations	23
# facts	192,532

- ▶ Closed-world
- ▶ Sparsity controllable

Metrics

Focus-on-top		***Less focus-on-top***
MRR	*vs*	*log-MRR*
Hits@1, 3, 10		*p-MRR*

Goal

- ▶ *Metric degradation & inconsistency w.r.t sparsity*
- ▶ *Effect of new metric*
- ▶ *Under the dependency assumption between β & ℓ*

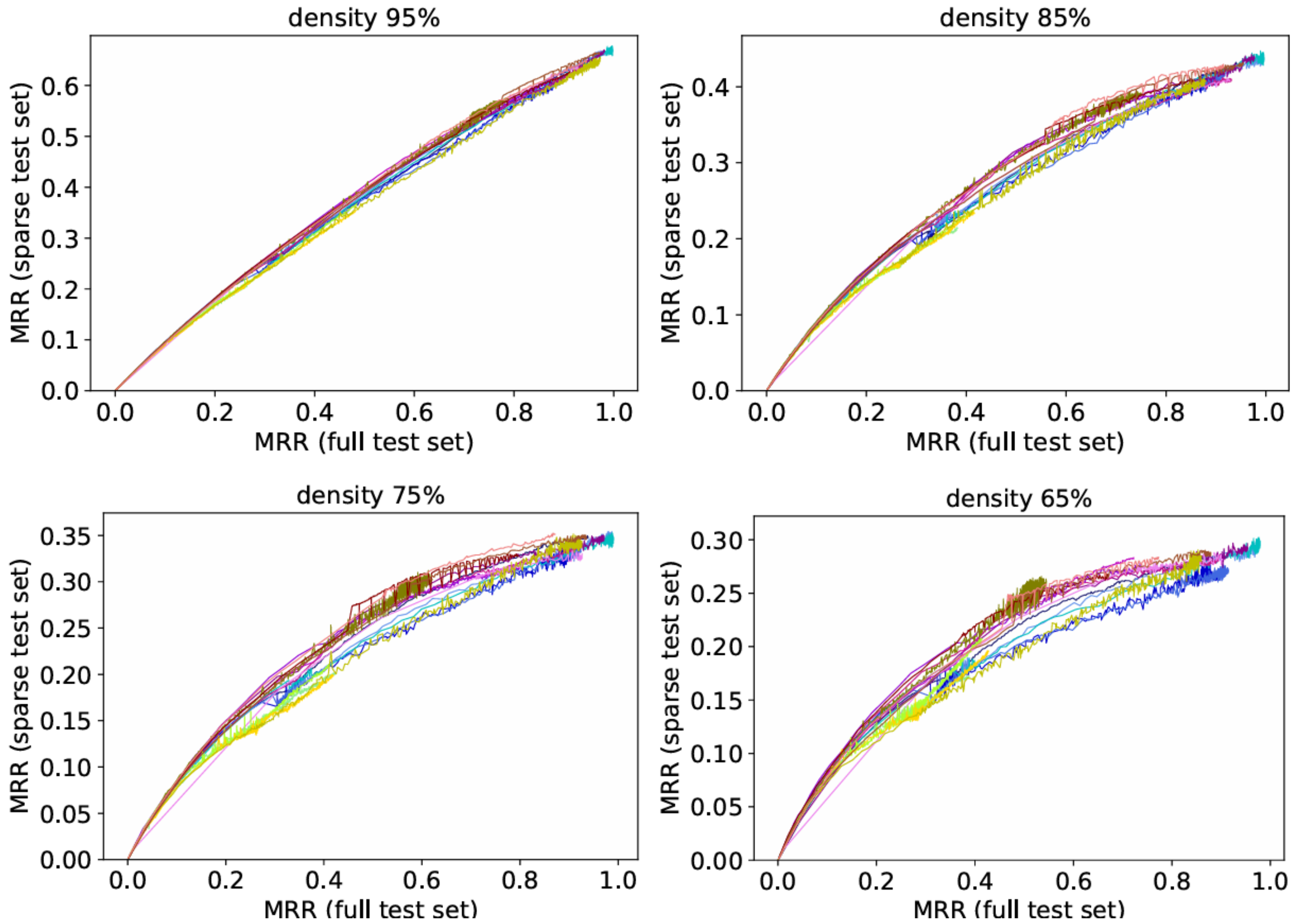
Models

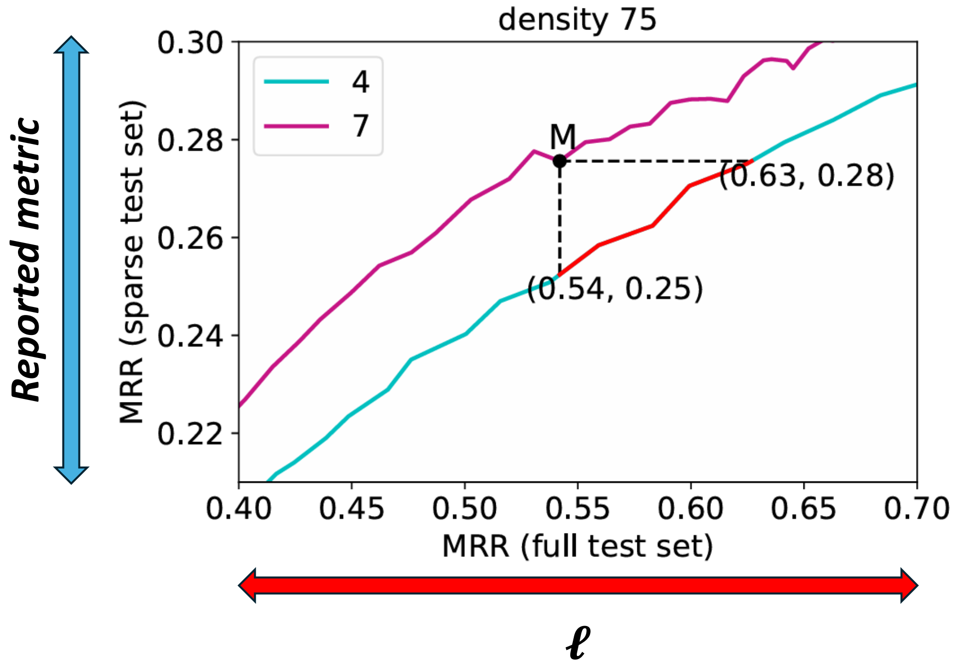
**different hyper-parameter settings*

- *RotatE*: 0, 1, 2, 3, 4
- *pRotatE*: 5, 6, 7, 8, 9
- *BetaE*: 10, 11, 12, 13, 14
- *ComplEx*: 15, 16, 17

EXPERIMENTS

■ MRR Degradation & Inconsistency (Independence Assumption)





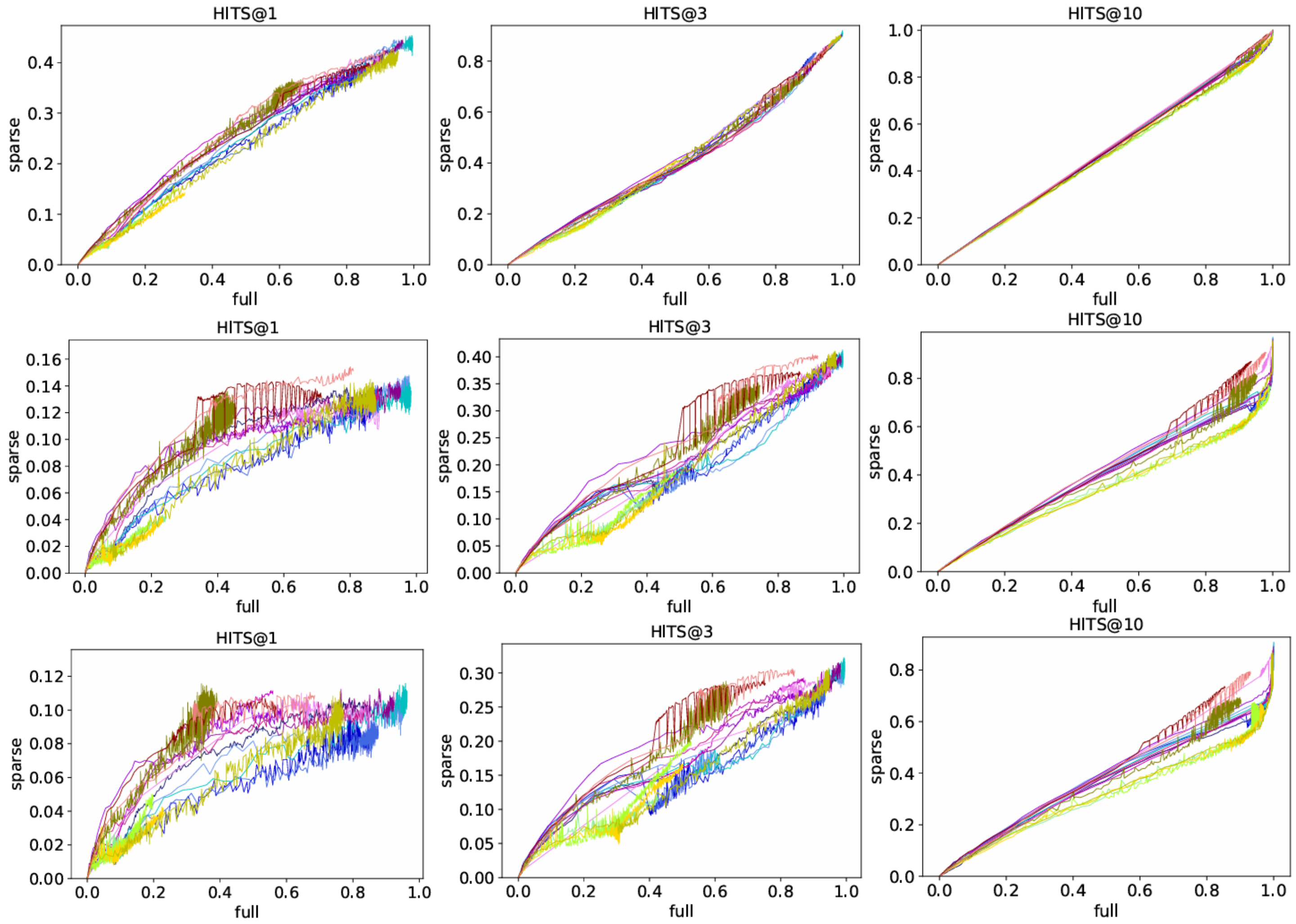
EXPERIMENTS

■ Hits@k Degradation & Inconsistency (Independence Assumption)

d=95%

d=75%

d=65%



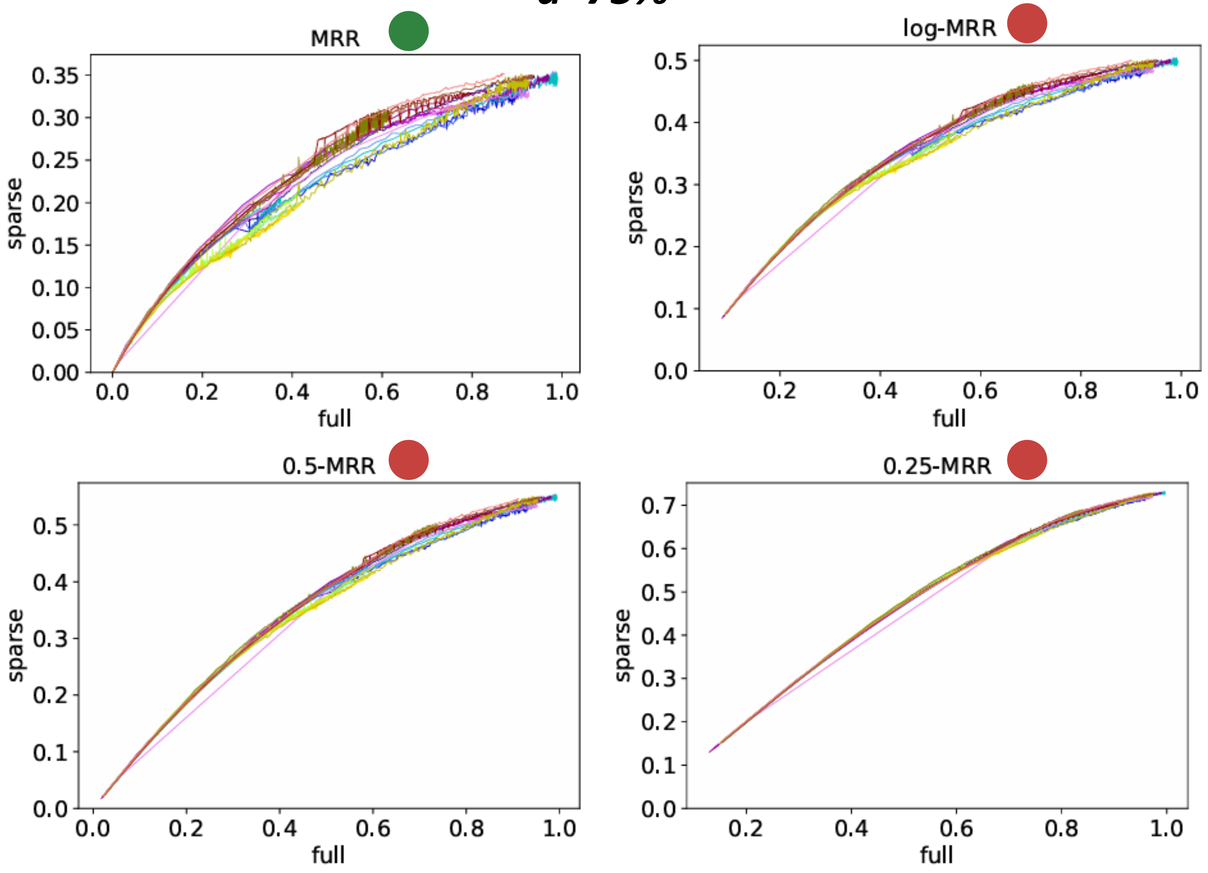
EXPERIMENTS

■ Alleviating Degradation and Inconsistency (Independence Assumption)

Focus-on-top vs *Less focus-on-top*

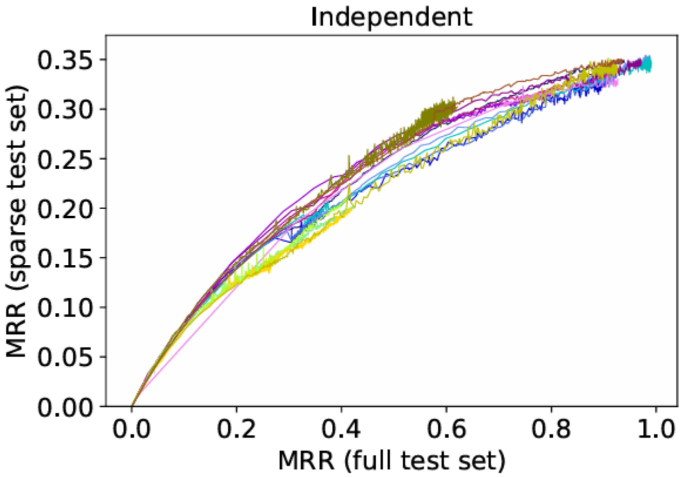
More consistent results acquired!

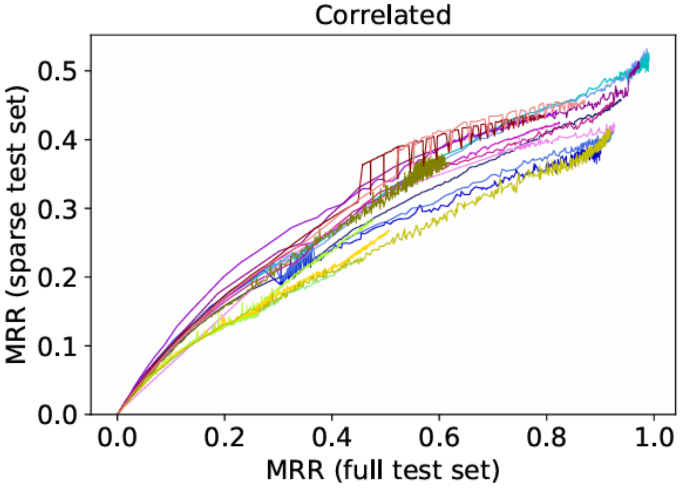
d=75%



EXPERIMENTS

Presence of Correlation (Dependence Assumption)





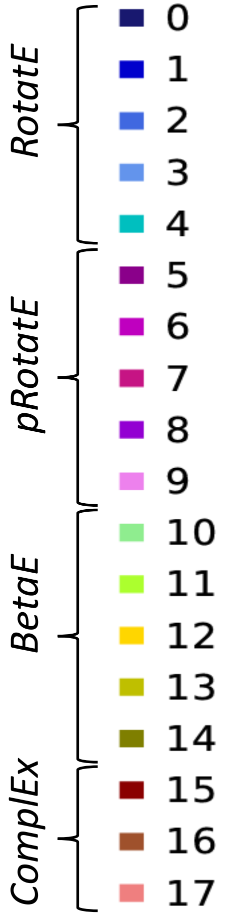
Complement sparse test set via **model 16** $\rho(\beta, \ell) < 0$

"MRR favor models with **smaller $\boldsymbol{\rho}$ instead of larger $\boldsymbol{\ell}$**"

Models correlated with **model 16** = $\rho(\beta, \ell) < 0$

e.g., **model 15** & **model 17**

Remove **model 16**

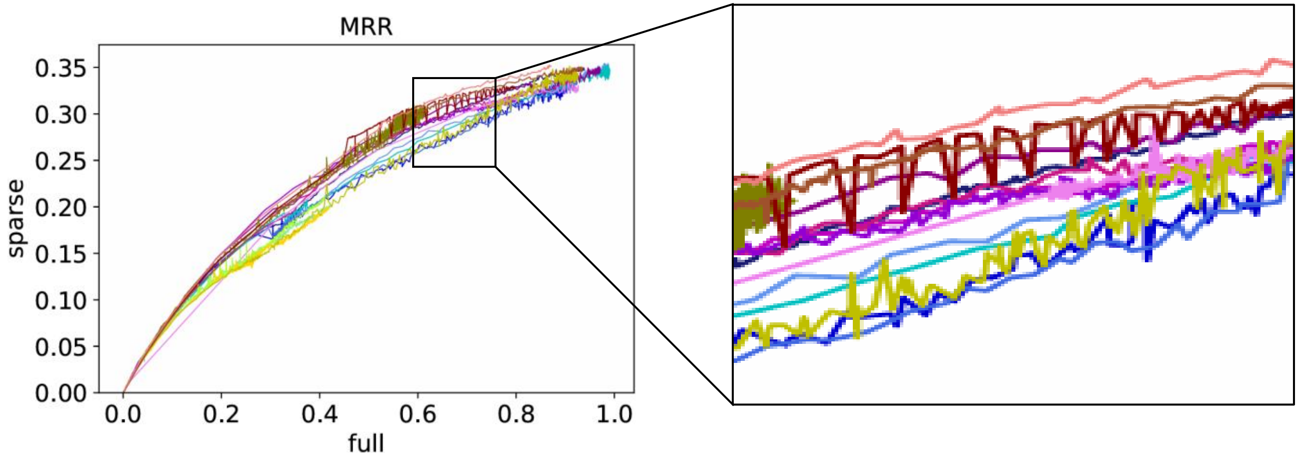


DISCUSSION

■ Illusion of Less Focus-on-Top?

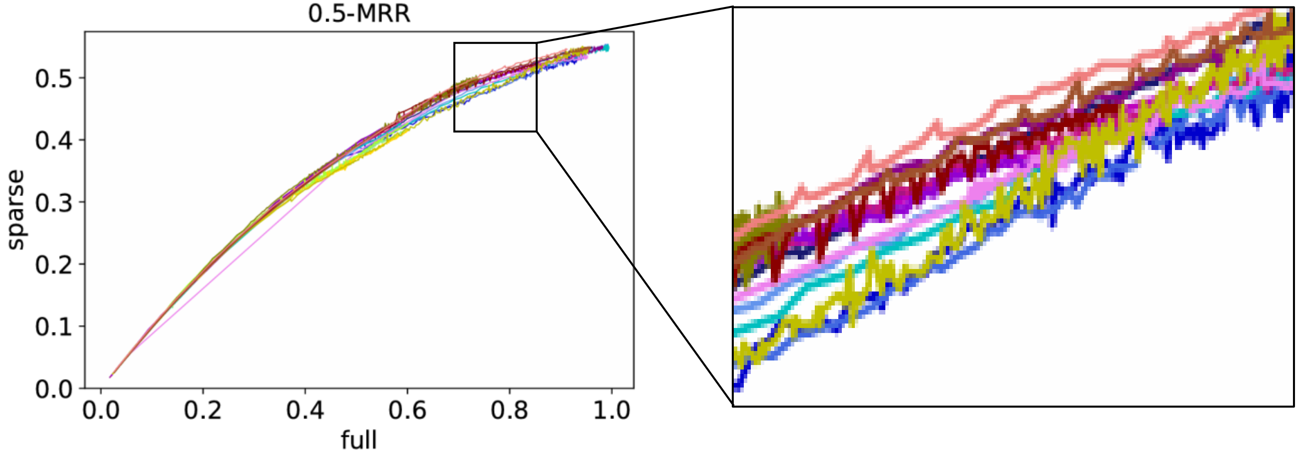
☐ Degradation problem is partially solved / is the inconsistency really alleviated?

Focus-on-top



Squeezed values of LFOT
→ smaller difference
→ illusion of consistency

Less focus-on-top

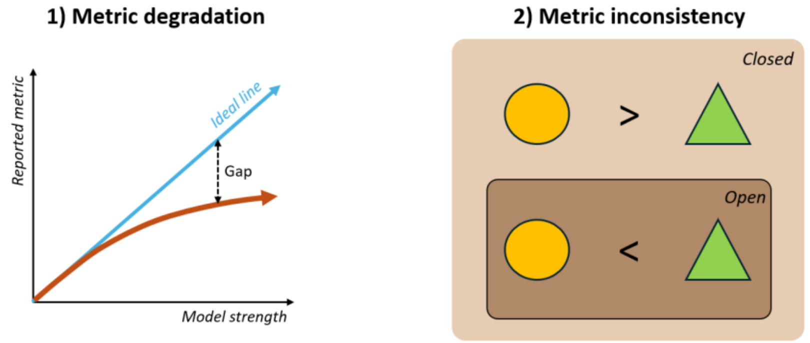


The most ideal metric in terms of
less degradation & consistency = $1/r^p (p \to 0)$
→ values collapse closer to each other

CONCLUSION

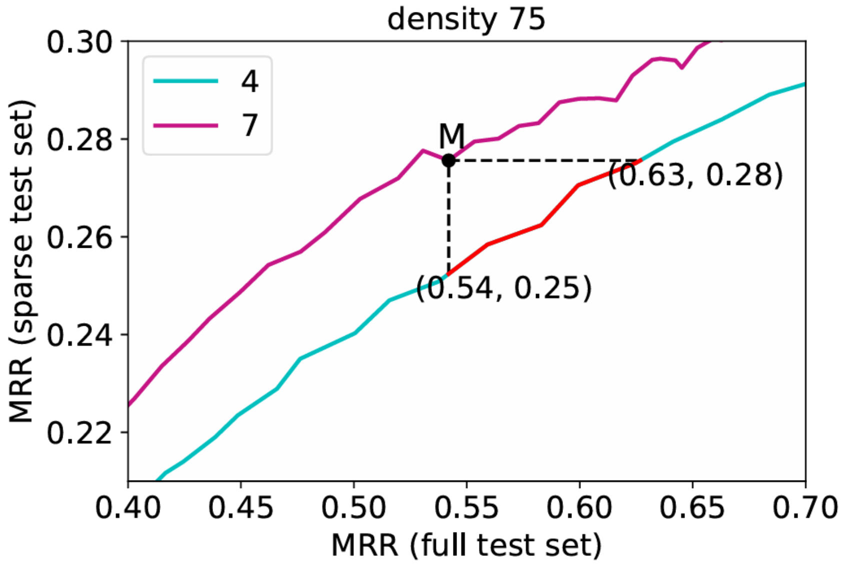
■ Discussion on Innate Trate of Open World Graph

- ☐ Theoretical investigation
- ☐ Metric's interpretation on model strength given sparsity



■ Metric Degradation & Inconsistency

- ☐ Acknowledging limitations of conventional metrics
- ☐ Proved via experiments



■ New Metrics

- ☐ Proposing less focus-on-top metrics
- ☐ Validated experimentally