

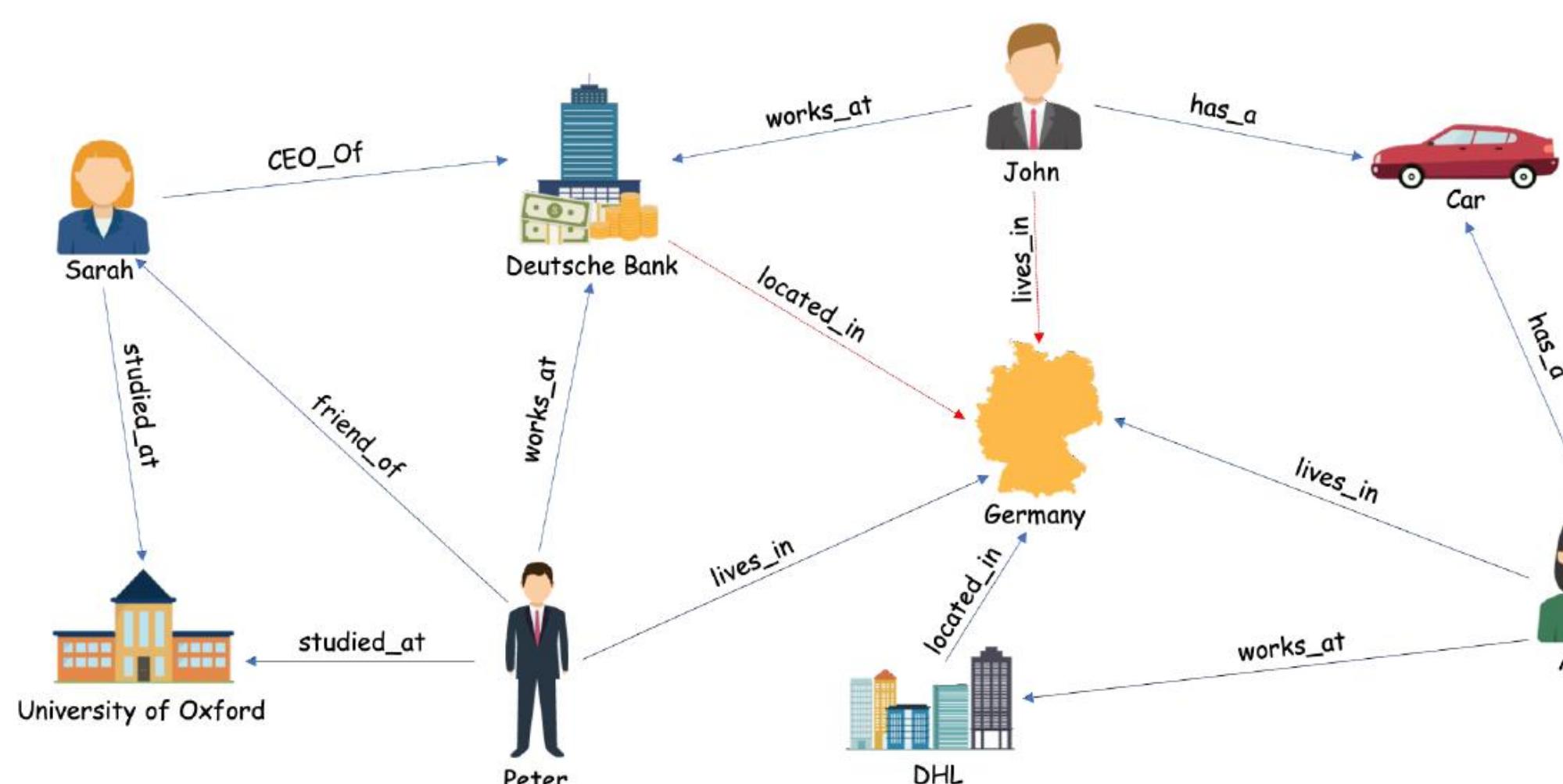
METRICHE PROPRIE E FLESSIBILI PER LA MISURAZIONE DELL'ACCURATEZZA NEI PROBLEMI DI LINK PREDICTION PER KNOWLEDGE GRAPH

LORENZO BALZOTTI – Università «La Sapienza» Roma lorenzo.balzotti@uniroma1.it | DONATELLA FIRMANI – Università «La Sapienza» Roma donatella.firmani@uniroma1.it |

GIOVANNA JONA LASINIO – Università «La Sapienza» Roma giovanna.jona-lasinio@uniroma1.it | MARCO LATTANZIO – Istituto Nazionale di Statistica lattanzio@istat.it |

KNOWLEDGE GRAPH (KG)

Basi dati di variabili categoriche che traducono in informazione connessioni reali tra entità (persone, luoghi, etc).



Connessioni in forma categorica sono tradotte in punti nello spazio vincolati tra loro tramite metriche

ETICHETTE			EMBEDDINGS		
TESTA	RELAZIONE	CODA	Head (h)	Relation (r)	Tail (t)
Sarah	CEO_di	Deutsche Bank	0.51364	-6.7586	11.57
John	Lives_in	Germany	1.116	4.45	-2.01
Anna	Works_at	DHL	3.21	9.9	0
...	...	...	...	...	...

FUNZIONE DI SCORING
 $\varphi(h, r, t) = s$

Padded scores

E' possibile applicare scoring rule **proprie** ad una trasformazione della distribuzione predittiva, nell' ottica di focalizzare l'attenzione sulle entità più probabili. Esempio: il Log padded score è definito come $-\ln(\hat{p}_t)$, dove $\hat{p}_t$ è una opportuna trasformazione della predittiva. (cfr. Resin 2023)

Softmax $s_t \rightarrow p_t$ Sparse softmax $p_t \rightarrow \hat{p}_t$ Scoring rule $\hat{p}_t \rightarrow -\ln(\hat{p}_t)$

Proper scoring rules

Le scoring rule $S(p, t)$ sono funzioni che danno un punteggio alla distribuzione predittiva al realizzarsi dell'entità vera. Come le **dart board nel lancio delle freccette**, sono **proprie** nel momento in cui il punteggio cresce quando la predittiva è «vicina» alla distribuzione vera, **unico caso** in cui lo score assume valore massimo. Le scoring rule possono essere viste come *loss* una volta che sono orientate negativamente. (cfr. Gneiting et al. 2007)

Approccio probabilistico

Il problema di link prediction si vede come un problema di classificazione probabilistica, trasformando score in probabilità.

$p(t|h, r) \longleftrightarrow q(t|h, r)$

Si misura la capacità della distribuzione predittiva p di riprodurre la distribuzione vera q tramite opportune ipotesi.

Obiettivo: p deve essere vicina a q .

Metriche non probabilistiche

Hits@k: % di triple del test set in cui l'entità vera è *suggerita* dal modello tra le prime k plausibili delle m totali possibili.

MRR: mean reciprocal rank, media aritmetica del reciproco del rank in cui l'entità esatta è suggerita dal modello
Cfr. Firmani et al. 2021



Metriche Padded scores sono strettamente proprie, Hits@k no

Il principale risultato del lavoro, sia in termini matematici che analitici attraverso degli esperimenti, è stato verificare che le metriche Hits@k non sono strettamente proprie e quindi, non discriminano tra modelli che, pur suggerendo in alta classifica le entità vere, lo fanno con probabilità diversa. I risultati preliminari sono stati ottenuti attraverso una *15-fold cross validation*, e calcolando gli intervalli di Tchebishev con valore $\lambda = 3$ (89% di probabilità di contenere il valore vero)

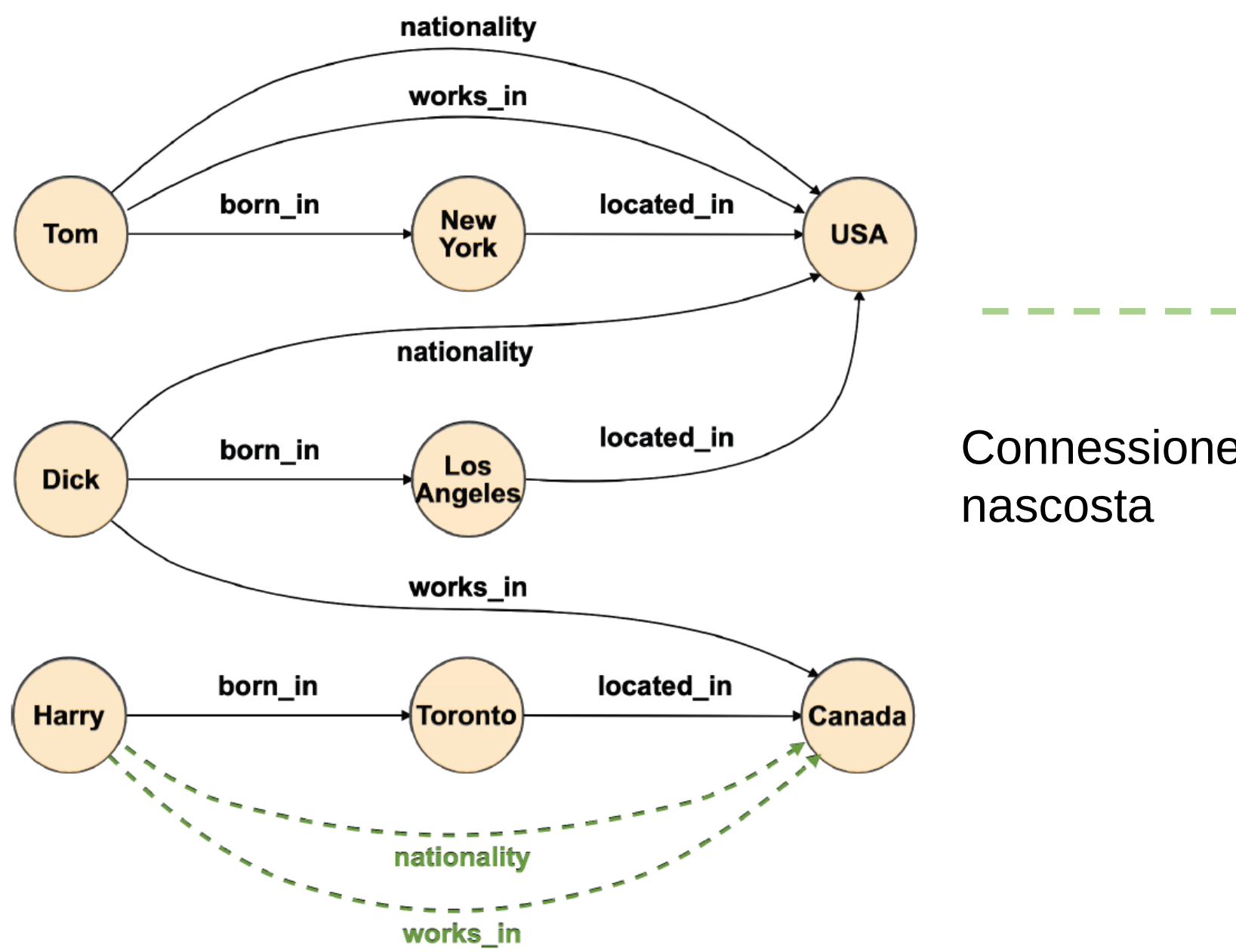
Risultati preliminari

Alcune evidenze: confrontando **Figura 1 e 2**, al crescere di K , le metriche Log scores sembrano stabilizzarsi rispetto alle metriche Hits, evidenziando un diverso comportamento nel valutare le probabilità rispetto al rank.

In **Figura 3** invece si evidenzia come focalizzando l'attenzione sulle prime 10 entità suggerite dal modello ($k=10$), le metriche Log score identificano il modello Rotate_CEL come best al pari delle metriche Hits, ma non identificano come Overlapping il modello TransE, diversamente dalla seconde.

LINK PREDICTION

Problema della ricerca di connessioni nascoste tra i dati. Machine Learning -> Predizione.



LEARNING SUI DATI.

UTILIZZO DI FUNZIONI LOSS E ALGORITMO GRADIENT DESCENT

Una volta deciso il modello di interazione tra gli *Embeddings*, la *funzione di scoring* φ e la *funzione loss*, si cerca l'insieme di score che minimizza la loss media attraverso il noto algoritmo *Gradient Descent*. Di solito si utilizza un sottoinsieme del KG definito dataset di *training*.

FUNZIONI LOSS PROBABILISTICHE E NON

Oltre ai tantissimi modelli di interazione studiati, è possibile scegliere diverse funzioni loss. Nei nostri esperimenti abbiamo scelto una loss probabilistica (*Cross Entropy Loss- CEL*) e una loss non probabilistica (*Margin Ranking Loss - MRL*).

ACCURATEZZA

La procedura di learning applicata sul training set viene utilizzata per *imparare dai dati*. La qualità del modello di *link prediction* così definito si valuta in base alla capacità di riprodurre connessioni note. In questa fase si utilizza il dataset di *test*.

Approccio non probabilistico

Voglio predire l'entità vera t che completa la tripla una volta note h ed r .

Si misura la capacità del modello di identificare come maggiormente plausibile l'entità vera t rispetto alle false.

Obiettivo: il *rank* r_t dell'entità vera deve essere alto.

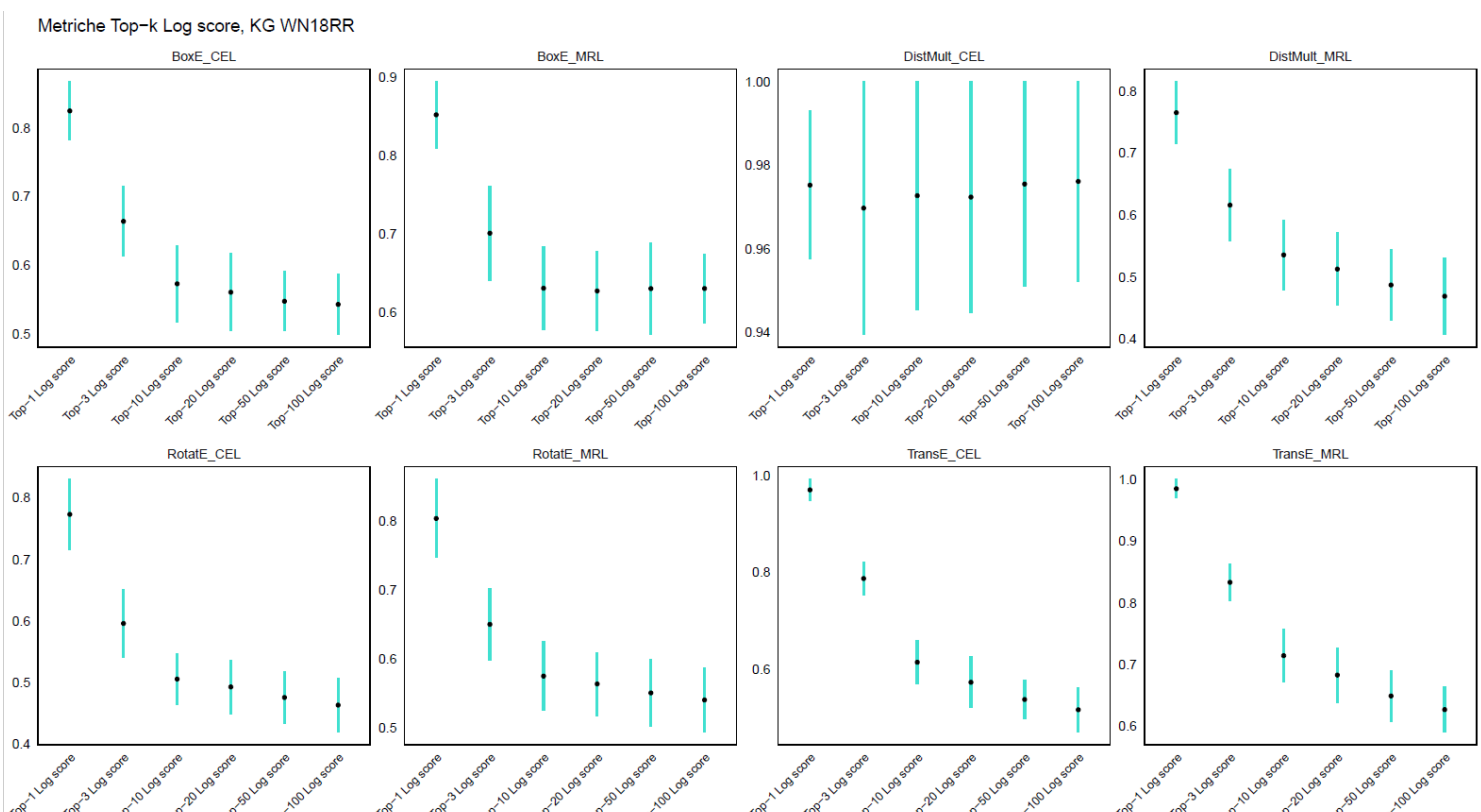


Figura 1: Nuove metriche Top-k Log score al crescere di k , Per diversi modelli, KG WN18RR

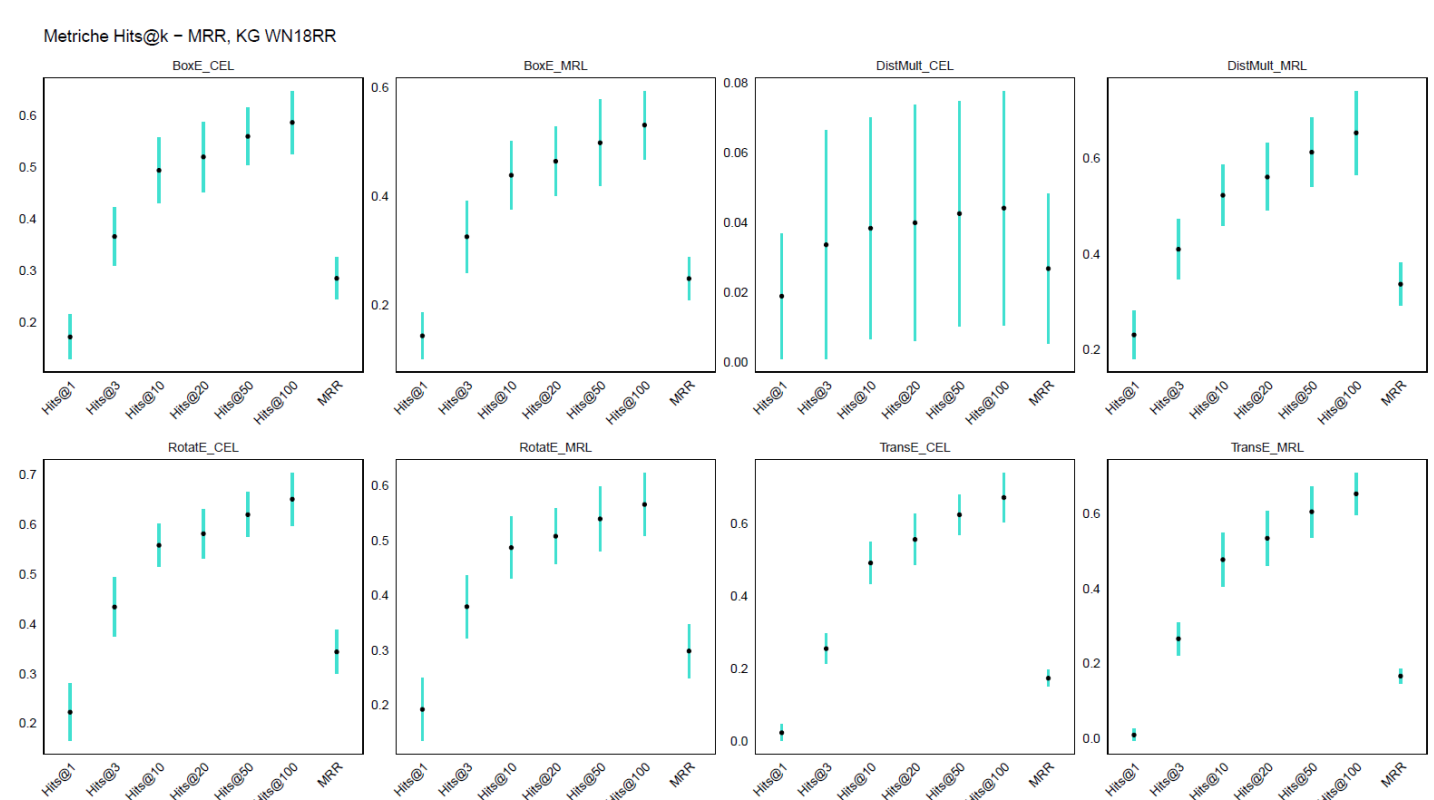


Figura 2: metriche Hits@k al crescere di k ed MRR, Per diversi modelli, KG WN18RR

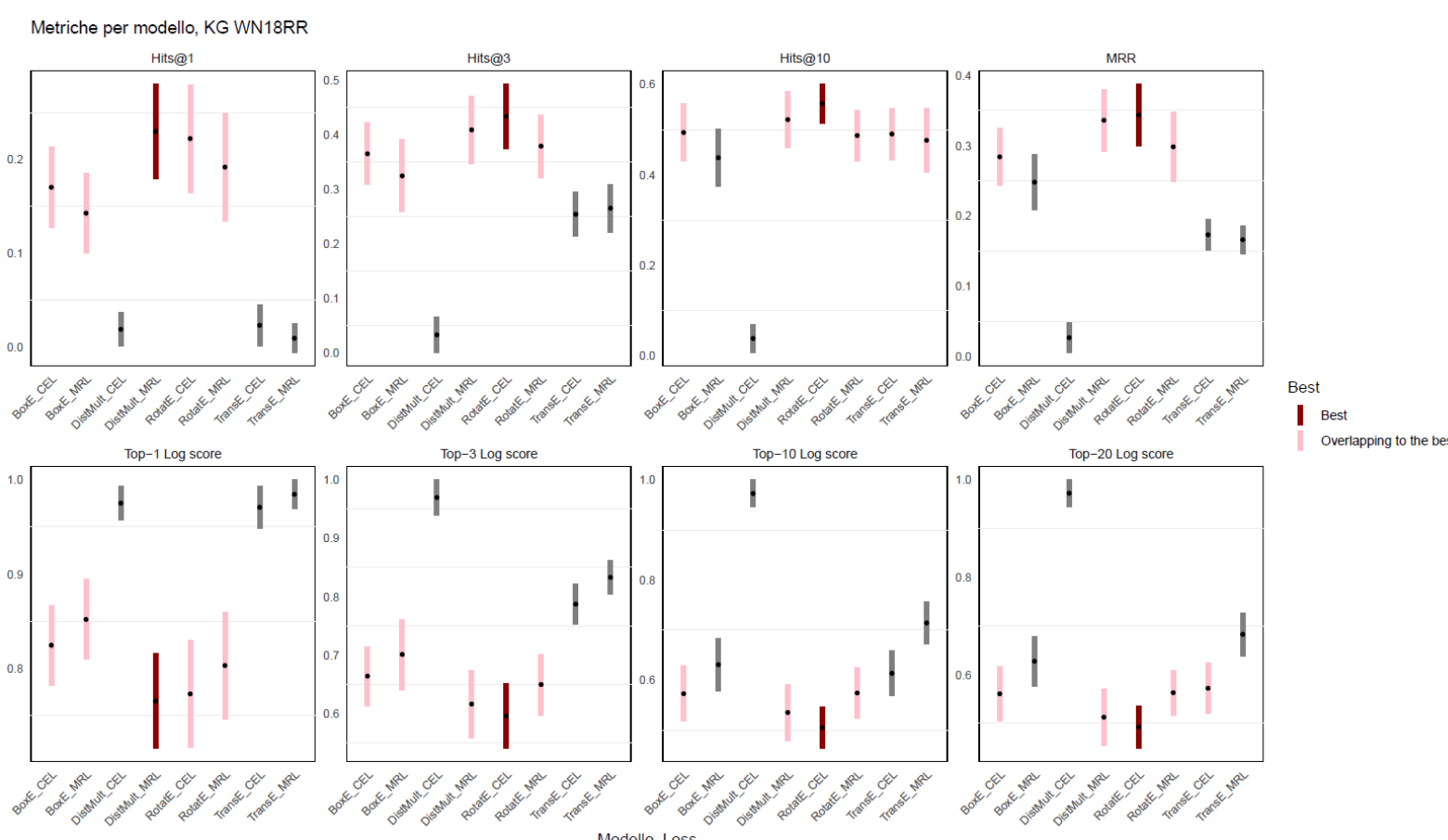


Figura 3: Confronto tra modelli per $k=1,3,10,20$, Metriche Hits@k, Top-k Log score, MRR, KG WN18RR

Bibliografia:
Gneiting et al., Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.
Firmani et al., Knowledge graph embedding for link prediction: A comparative analysis. ACM Transactions on Knowledge Discovery from Data, 15(2):1–49, January 2021.
Resin, From classification accuracy to proper scoring rules: Elicitability of probabilistic top list predictions. Journal of Machine Learning Research, 24(173):1–21, 2023.