Evaluating Cumulative Spectral Gradient as a Complexity Measure

School of Digital Science, Universiti Brunei Darussalam, Brunei Darussalam
Haji Gul (23h1710@ubd.edu.bn), Abdul Ghani Naim, Ajaz Ahmad Bhat (ajaz.bhat@ubd.edu.bn)

Introduction

- This study evaluates the complexity of the tail prediction task in knowledge graphs.
- Challenge: Metrics like MRR measure performance but not dataset complexity.

Research Questions:

- Is CSG sensitive to parameters K (neighbors) and M (samples)?
- Does CSG correlate with MRR in KGs?
- How do K and M (sample size) influence CSGs' complexity estimation?

Methodology

Step 1) Grouping by Tail Entities:

$$T = \{(h_i, r_i, t_i) \mid h_i \in E, r_i \in R, t_i \in E\},$$

$$G(C_i) = \{(h, r) \mid (h, r, C_i) \in T\}, \quad \forall C_i \in E,$$

resulting in a mapping:

$$C_i \mapsto G(C_i),$$

$$C_i = \{C_1, C_2, \ldots, C_K\},$$

Step 2) Generate embeddings:

$$e_h = \text{BERT}(h) \in \mathbb{R}^d, \quad e_r = \text{BERT}(r) \in \mathbb{R}^d,$$

$$\phi(h, r) = e_h \oplus e_r \in \mathbb{R}^{2d},$$

$$\Phi(C_i) = \{\phi(h, r) \mid (h, r, C_i) \in T\},$$

Step 3) Build a similarity matrix:

$$S_{ij} = \frac{1}{Mk} \sum_{\phi_m \in \Phi(C_i)_{\text{sample}}} \sum_{\phi_n \in K(\phi_m)} \mathbb{I}[\phi_n \in \Phi(C_j)],$$

where the indicator function is:

$$\mathbb{I}[\phi_n \in \Phi(C_j)] = \begin{cases} 1, & \text{if } \phi_n \in \Phi(C_j), \\ 0, & \text{otherwise.} \end{cases}$$

Step 4) Graph Laplacian and Spectral Analysis:

$$D_{ii} = \sum_{j=1}^{K} S_{ij}, \quad D_{ij} = 0 \text{ for } i \neq j.$$

$$L = I - D^{-1/2} S D^{-1/2},$$

$$D_{ii}^{-1/2} = \frac{1}{\sqrt{D_{ii}}}, \quad \text{for } D_{ii} > 0.$$

$$L u_i = \lambda_i u_i, \quad u_i \in \mathbb{R}^K, \quad \|u_i\| = 1, \quad 0 \leq \lambda_i \leq 2.$$

Step 5) Cumulative Spectral Gradient:

$$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{K-1},$$

$$\text{Define gaps, } \delta_i = \lambda_{i+1} - \lambda_i, \quad i = 0, 1, \ldots, K-2,$$

$$\text{Then, } \text{CSG}_{k_c} = \sum_{i=0}^{k_c - 1} \delta_i = \lambda_{k_c} - \lambda_0,$$

$$\text{and, } \text{CSG}_{K-1} = \lambda_{K-1} - \lambda_0.$$

Methodology Graphical View:

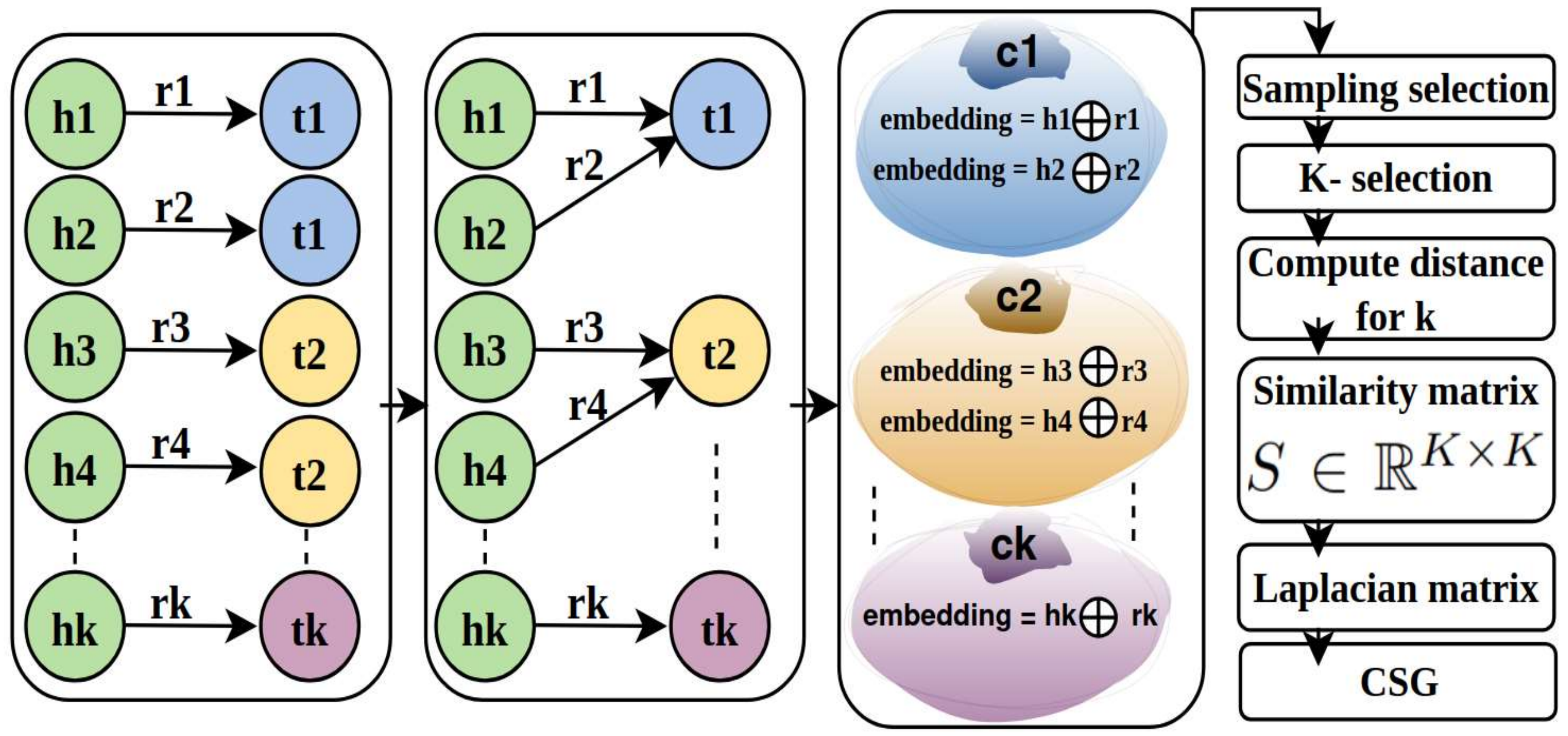

Figure 1: ***Proposed KG-CSG Methodology***

Results

- ***Datasets***: FB15k-237, WN18RR, CoDEx-S, CoDEx-M, CoDEx-L
- ***Sensitivity to M:*** For small K, M become stable CSG, but its impact is less pronounced than K's.

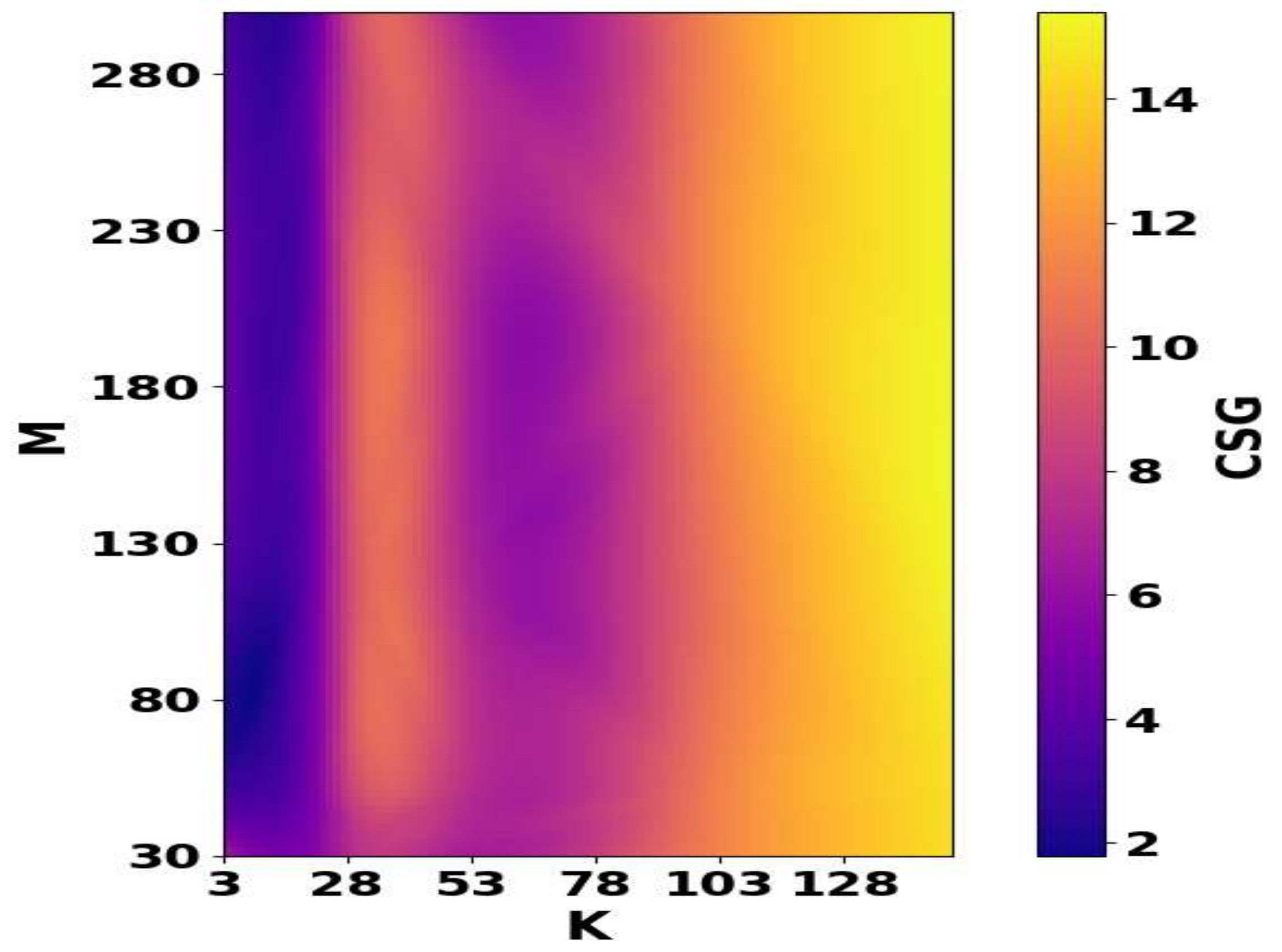

Figure 2: ***CSG as a function of M and K values***

- ***Role of K:*** CSG is deeply sensitive to K.

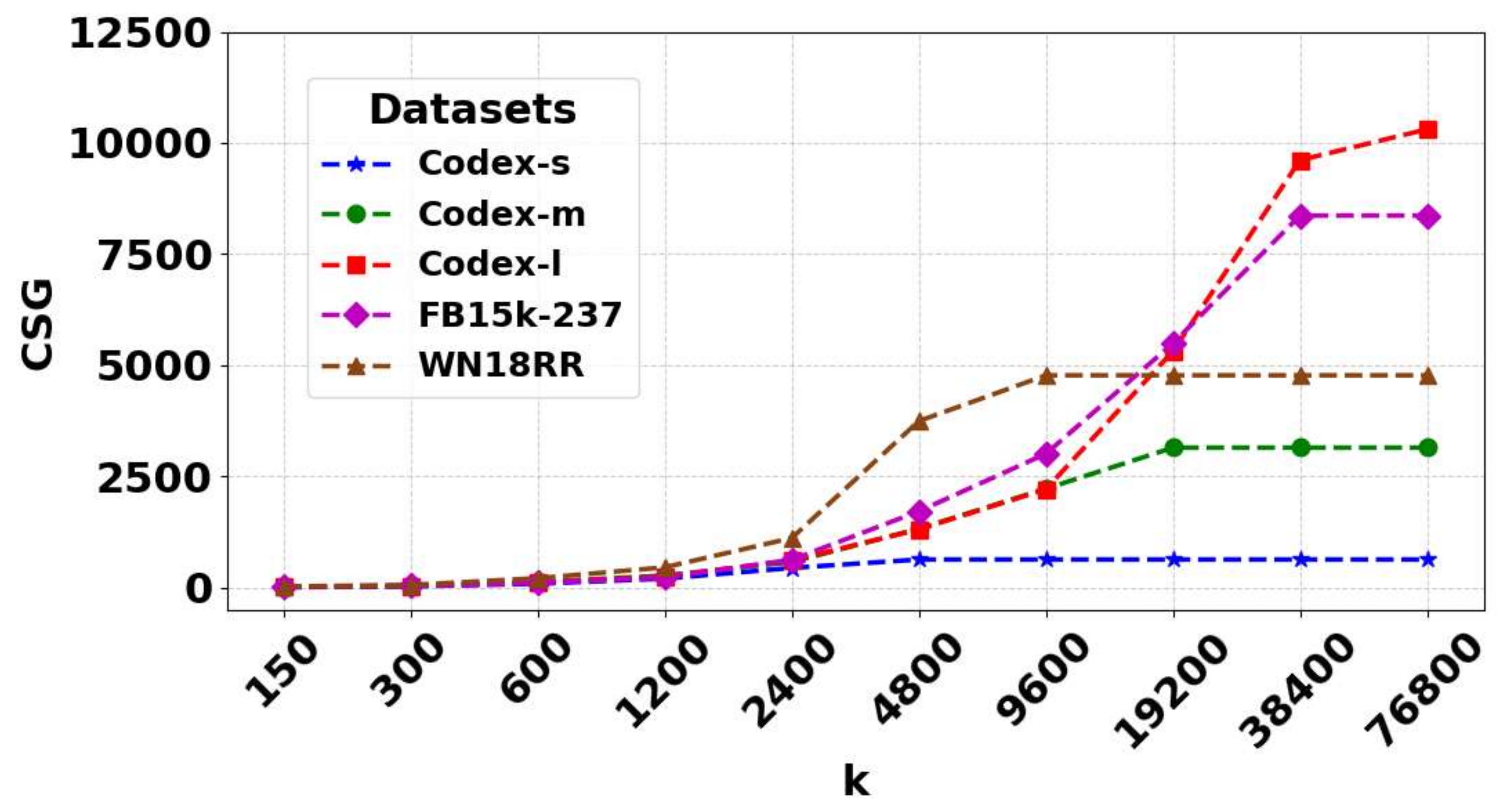

Figure 3: ***CSG as a function of K values at M = 100***

- ***Weak MRR Correlation:*** CSG does not show much strong correlation with ***MRR*** (***R = −0.64***).

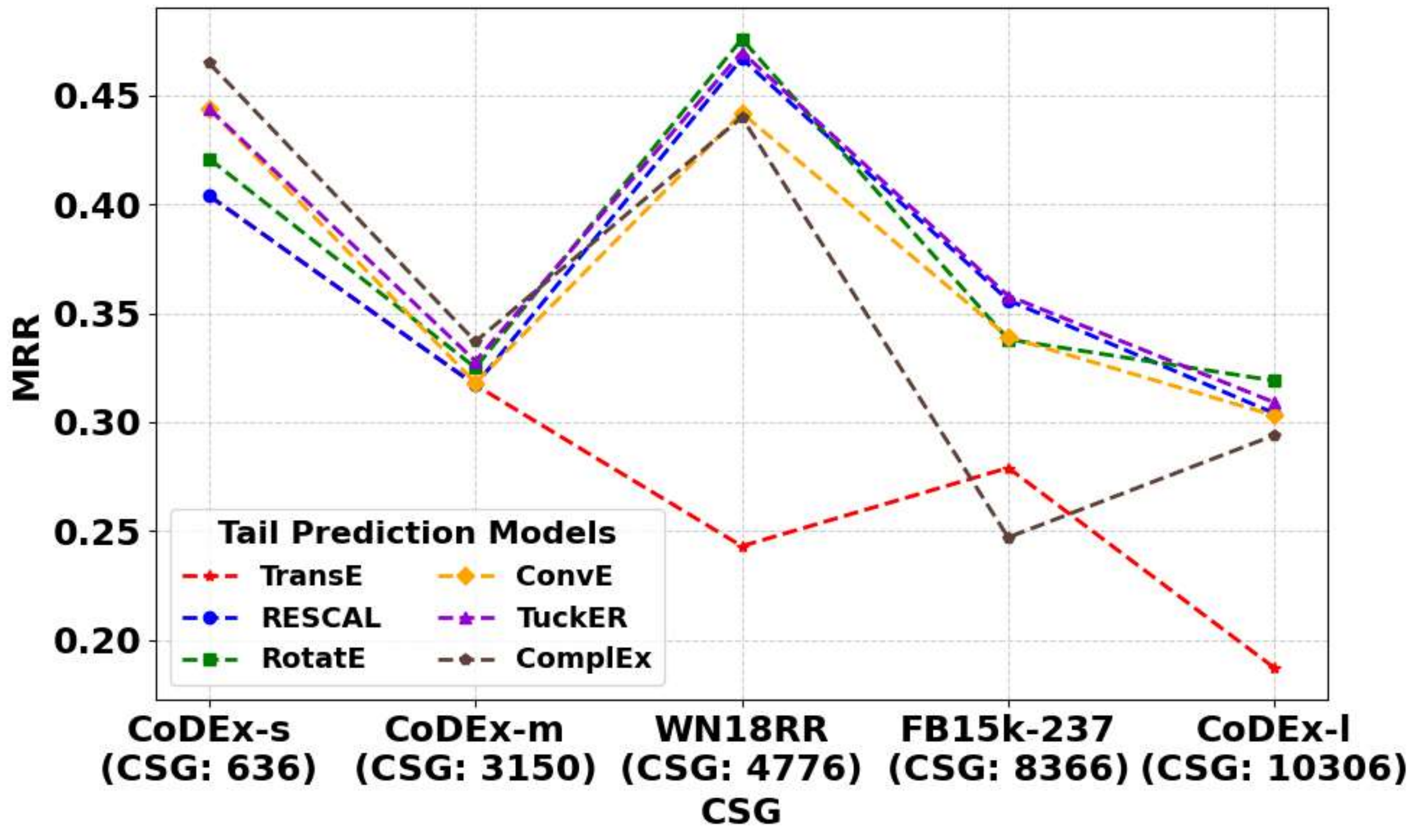

Figure 4: ***Relation between CSG and MRR.***

Conclusion

- ***CSG*** is significantly influenced by the ***K*** and ***M***, challenging previous assumptions that K and M had minimal impact
- Parameters K and M deeply influence results.
- Poor relation between ***CSG*** and performance (***MRR***).
- Future work focus on developing complexity measures tailored to the characteristics of knowledge graphs.