# Evaluating Cumulative Spectral Gradient as a Complexity Measure

**Haji Gul** [1] **Abdul Ghani Naim** [1] **Ajaz Ahmad Bhat**✉ [1]

,

## Abstract

Accurate estimation of dataset complexity is crucial for evaluating and comparing link-prediction models for knowledge graphs (KGs). The Cumulative Spectral Gradient (CSG) metric (Branchaud-Charron et al., 2019) —derived from probabilistic divergence between classes within a spectral clustering framework— was proposed as a dataset complexity measure that (1) naturally scales with the number of classes and (2) correlates strongly with downstream classification performance. In this work, we rigorously assess CSG's behavior on standard knowledge-graph link-prediction benchmarks—a multi-class tail-prediction task— using two key parameters governing its computation: $M$, the number of Monte Carlo–sampled points per class, and $K$, the number of nearest neighbors in the embedding space. Contrary to the original claims, we find that (1) CSG is highly sensitive to the choice of $K$, thereby does not inherently scale with the number of target classes, and (2) CSG values exhibit weak or no correlation with established performance metrics such as mean reciprocal rank (MRR). Through experiments on FB15k-237, WN18RR, and other standard datasets, we demonstrate that CSG's purported stability and generalization-predictive power break down in link-prediction settings. Our results highlight the need for more robust, classifier-agnostic complexity measures in KG link-prediction evaluation.

## 1. Introduction

Knowledge graphs (KGs) underlie many high-impact applications—ranging from recommendation systems (Spillo et al., 2024) and question answering (Zeng et al., 2025) to drug discovery (Zhang et al., 2025; Gul et al., 2025a). By encoding relational knowledge as triples $(h, r, t)$, link prediction $(h, ?, t)$ and entity prediction $(h, r, ?)$ tasks on KGs enable models to infer missing relations or tail entities (Gul et al., 2024; 2025b). These benchmarks however, remain challenging due to imbalanced class distributions and overlapping feature patterns across relations and entities (Bourli & Pitoura, 2020). While metrics like MRR and Hits@k evaluate how accurately models retrieve correct links, these do not provide us a direct measure of the intrinsic complexity of KG datasets under various link prediction scenarios. A robust class-separability measure would (a) quantify dataset complexity across different link-prediction formulations—revealing, for example, whether predicting rare drug-target pairs is inherently harder than predicting common entity relations— (b) anticipate generalization performance—setting realistic expectations for new methods before expensive downstream evaluation, and (c) facilitate a unified estimate of model performance across datasets.

CSG is a recently proposed spectral metric—derived from the eigenvalues of the normalized graph Laplacian—designed to quantify dataset complexity by measuring class separability. In image classification benchmarks, higher CSG values correlate strongly with lower test accuracy (Branchaud-Charron et al., 2019). However, KG link-prediction classification is a large-scale, multi-class task with thousands of candidate tails (every graph entity). In this regime, two core claims of CSG merit re-evaluation:

- CSG's reliance on the nearest-neighbor parameter $K$ may prevent it from naturally scaling when the number of target classes grows to typical KG sizes.
- Although CSG correlates with accuracy in image tasks, it is unknown whether CSG scores—computed over embeddings from KG models (e.g., BERT-based or translational)—correlate with standard KG metrics like Mean Reciprocal Rank (MRR).

No prior work has systematically evaluated CSG on canonical KG benchmarks (e.g., FB15k-237, WN18RR) or examined its sensitivity to the Monte Carlo sample size $M$ and neighbor count $K$. Empirical scrutiny is therefore required

*Equal contribution [1]School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam. Correspondence to: Ajaz Ahmad Bhat <ajaz.bhat@ubd.edu.bn>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

to assess CSG's stability and predictive power in large-scale link-prediction settings.

To this end, we conduct the **first systematic evaluation of CSG in multi-class tail-prediction tasks across multiple standard KG datasets** (e.g., FB15k-237, WN18RR). Specifically for each head–relation pair, we treat every candidate tail entity as a separate class. We vary the Monte Carlo sample size $M$ and the nearest-neighbor count $K$ to compute and relate CSG values over embeddings. Further on, we compare CSG values against actual link-prediction performance (MRR) to quantify how well CSG predicts generalization. We list below key findings of our work:

1. **Sensitivity to $K$:** CSG values change dramatically as $K$ varies, showing that any perceived "scalability" with the number of classes is an artifact of specific dataset choices rather than an inherent property of CSG.
2. **Weak Performance Correlation:** Across all datasets and KG models, CSG scores exhibit near-zero Pearson correlation with MRR, contradicting the claim that CSG reliably predicts downstream accuracy.

These results question CSG's utility as a model-agnostic separability metric for large-scale classification and highlight the need for more robust measures in KG evaluation.

## 2. Methodology

In our approach for CSG computation, we transform KG triplets into multi-class representations, use BERT embeddings for semantic richness, and apply spectral analysis to derive the CSG values; see Figure 1 for more clarification. The following subsections detail each step of this process.

***Grouping by Tail Entities:*** Knowledge graphs, such as FB15k-237 and WN18RR, consist of a set of triplets:

$$T = \{(h_i, r_i, t_i) \mid h_i \in E, r_i \in R, t_i \in E\}, \quad (1)$$

where $h_i$ is the head entity, $r_i$ is the relation, and $t_i$ is the tail entity, with $E$ being the set of all entities and $R$ the set of all relations. The next step organizes this data by grouping triplets according to their tail entities class $C$ (*each unique* $t_i \rightarrow$ *denotes a unique class* $C_i$) using a mapping function:

$$G(C_i) = \{(h, r) \mid (h, r, C_i) \in T\}, \quad \forall C_i \in E, \quad (2)$$

resulting in a mapping:

$$C_i \mapsto G(C_i), \quad (3)$$

which aggregates all $(h, r)$ pairs pointing to the same tail $C_i$. Each unique tail entity is treated as a distinct class, forming a set:

$$C_i = \{C_1, C_2, \ldots, C_K\}, \quad (4)$$

$K$ is the total count of unique tails, generating $K$ classes based on tail entities for further examination.

**Generating Embeddings:** To transform textual head entities and relations into numerical form, a pre-trained BERT model generates dense vector embeddings. For each head entity $h$ and relation $r$, embeddings are:

$$e_h = \text{BERT}(h) \in \mathbb{R}^d, \quad e_r = \text{BERT}(r) \in \mathbb{R}^d, \quad (5)$$

$d$ the embedding dimension is. BERT-Base (Hugging Face Transformers) was used to generate 768-dimensional embeddings, preprocessing head entities and relations as single tokens. For every triplet $(h, r, C_i) \in T$, a composite vector is formed:

$$\phi(h, r) = e_h \oplus e_r \in \mathbb{R}^{2d}, \quad (6)$$

where $\oplus$ denotes concatenation. These composite vectors are then grouped according to their corresponding tail entities:

$$\Phi(C_i) = \{\phi(h, r) \mid (h, r, C_i) \in T\}, \quad (7)$$

each tail $C_i$ is associated with a set of $(h, r)$ vectors. This step provides a meaningful representation of the triplet data, organized by tail classes, preparing the data for complexity analysis.

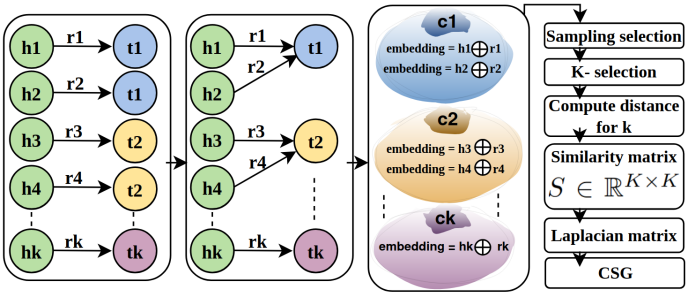


*Figure 1.* Left box showing triplets where the heads are $(h_1, h_2, \ldots, h_k)$ green, relations $(r_1, r_2, \ldots, r_k)$, tails $(t_1, t_2, t_3)$ are in blue, yellow and purple. The next box denotes the grouping of their tail entities into classes: $c_1$ for $t_1$, with $(h_1, r_1, t_1)$ and, $(h_2, r_2, t_1)$ belonging to the same class, for example. BERT is used to embed head-relation pairs, producing 768-dimensional vectors, and then concatenates them, such as $h_1 \oplus r_1$ and $h_2 \oplus r_2$, for class $c_i$. Next, a sampled $K$ search is performed to compute distances and a similarity matrix $S \in \mathbb{R}^{K \times K}$. The Laplacian matrix $L$ is obtained, and $S$ the spectral complexity of the KG is quantified using the CSG calculated from its eigenvalues.

**Similarity Computation and Matrix Construction:** A similarity matrix $S \in \mathbb{R}^{K \times K}$ is constructed, where $K$ is the number of classes. Let $\Phi(C_i)$ denote the set of vectors for class $C_i$. Each vector:

$$\phi_m = e_h \oplus e_r \in \mathbb{R}^{2d}. \quad (8)$$

A subset is sampled:

$$M = \min(N, |\Phi(C_i)|),$$
$$\Phi(C_i)_{\text{sample}} = \{\phi_1, \phi_2, \ldots, \phi_M\} \subset \Phi(C_i) \quad (9)$$

where $N$ is the number of vector samples per class. To manage computational complexity for large KGs, we sample $M = 120$ vectors per class. For each $\phi_m \in \Phi(C_i)_{\text{sample}}$, compute its $k = 50$-nearest neighbors via L2 distance:

$$\|\phi_m - \phi_n\|_2^2 = \sum_{l=1}^{2d} (\phi_{m,l} - \phi_{n,l})^2. \tag{10}$$

it computes the Euclidean distance between two concatenated BERT embeddings, $\boldsymbol{\phi}_m$ and $\boldsymbol{\phi}_n$, where $\boldsymbol{\phi}_m, \boldsymbol{\phi}_n \in \mathbb{R}^{2d}$ represent the combined head-relation embeddings of triplets, respectively. while $\phi_{m,l}$ and $\phi_{n,l}$ denote the $l$-th components of the respective vectors. This distance metric is used during $k$ values neighbor (k-NN) search to measure nearest neighbor triplets grouped by tail entities, enabling the construction of the class similarity matrix $S$, which can be defined as in Equation 11. The distance computation directly impacts the spectral analysis by indicating how tightly or loosely classes overlap, thereby influencing the Cumulative Spectral Gradient (CSG), a measure of dataset complexity derived from the eigenvalue gaps in the graph Laplacian.

$$S_{ij} = \frac{1}{Mk} \sum_{\phi_m \in \Phi(C_i)_{\text{sample}}} \sum_{\phi_n \in K(\phi_m)} \mathbb{I}[\phi_n \in \Phi(C_j)], \tag{11}$$

where the indicator function is:

$$\mathbb{I}[\phi_n \in \Phi(C_j)] = \begin{cases} 1, & \text{if } \phi_n \in \Phi(C_j), \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

$\phi_n$ is an embedding vector, $\Phi(C_j)$ denotes the set of embeddings for class $C_j$, and $I$ is an indicator function returning 1 if $\phi_n$ belongs to $C_j$. It is employed in the formation of the similarity matrix $S$ to enumerate the $K$-nearest neighbors of the class $C_i$ that belong to $C_j$, quantifying inter-class overlap for complexity analysis.

**Graph Laplacian and Spectral Analysis:** Graph Laplacian captures the connectivity and clustering tendencies of the classes, rooted in graph theory and spectral analysis. The normalized Laplacian provides a standardized measure of how classes are interconnected, accounting for variations in their degrees of connection. The graph Laplacian captures class connectivity and clustering tendencies. The diagonal degree matrix $D \in \mathbb{R}^{K \times K}$ can be defined as Equation 13, while the normalized Laplacian as Equation 14.

$$D_{ii} = \sum_{j=1}^{K} S_{ij}, \quad D_{ij} = 0 \text{ for } i \neq j. \tag{13}$$

$$L = I - D^{-1/2} S D^{-1/2}, \tag{14}$$

where $I$ is the $K \times K$ identity matrix, and:

$$D_{ii}^{-1/2} = \frac{1}{\sqrt{D_{ii}}}, \quad \text{for } D_{ii} > 0. \tag{15}$$

$D_{ii} = \sum_{j=1}^{K} S_{ij}$ representing the total similarity of a class $C_i$ to all other classes, where $D_{ii}^{-1/2} = \frac{1}{\sqrt{D_{ii}}}$, ensures eigenvalues. Compute eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{K-1}$ and eigenvectors $u_1, u_2, \ldots, u_K$ from:

$$Lu_i = \lambda_i u_i, \quad u_i \in \mathbb{R}^K, \quad \|u_i\| = 1, \quad 0 \leq \lambda_i \leq 2. \tag{16}$$

yields eigenvalues $\lambda_i$ and orthonormal eigenvectors $u_i$, which encode structural properties.

**Cumulative Spectral Gradient (CSG) Computation:** Defines a complexity measure based on the differences between consecutive eigenvalues of the Laplacian, summing them cumulatively to assess how the graph's structure evolves across its spectrum. Theoretically, the CSG quantifies the cumulative effect of spectral gaps, reflecting the progressive separation of classes and providing a nuanced view of complexity that ties directly to the graph's global properties. This is particularly relevant for tail prediction, as it indicates the degree of variation in prediction difficulty across the dataset. The CSG measures complexity via eigenvalue differences. Order the eigenvalues:

$$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{K-1}, \tag{17}$$

$$\text{Define gaps, } \delta_i = \lambda_{i+1} - \lambda_i, \quad i = 0, 1, \ldots, K-2, \tag{18}$$

$$\text{Then, } \text{CSG}_{k_c} = \sum_{i=0}^{k_c - 1} \delta_i = \lambda_{k_c} - \lambda_0, \tag{19}$$

$$\text{and, } \text{CSG}_{K-1} = \lambda_{K-1} - \lambda_0. \tag{20}$$

Branchaud-Charron et al. (2019) claim that higher CSG values indicate higher complexity (more class overlap); lower CSG values indicate better separation and easier tail prediction.

### 2.1. Experiments

**Datasets:** The following datasets are used: **FB15k-237** (Bollacker et al., 2008) consists of 14,541 entities, 237 relations, and a total of 310,116 triplets. **WN18RR** (Miller, 1995) includes 40,943 entities, 11 relations, and a total of 92,583 triplets. **CoDEx-S** (Safavi & Koutra, 2020) features 2,034 entities, 42 relations, and a total of 40,198 triplets. **CoDEx-M** (Safavi & Koutra, 2020) has 17,050 entities, 51 relations, and a total of 185,584 triplets. **CoDEx-L** (Safavi & Koutra, 2020) includes 77,951 entities, 69 relations, and a total of 673,872 triplets. **Countries** (Liang et al., 2024) dataset contains 271 entities, 2 relations, and 1159 triplets. **Toy** (Liang et al., 2024) includes 278 entities, 112 relations, and 4826 triplets. **UML** (Bodenreider, 2004) contains 135 entities, 46 relations, and 6529 triplets. **Nations** (Liang et al., 2024) has 14 entities, 55 relations, and 1992 triplets. All the datasets are publicly available online [1].

[1] https://tinyurl.com/mr8ckwmb

### 2.2. Results

Figure 2 illustrates the Cumulative Spectral Gradient (CSG) of Codex-S dataset in the tail-prediction task setting, represented as a surface function of parameters $K$ and $M$. Both parameters influence the CSG values, the impact of $K$ on CSG however is quite significant. This observation counters the previously held belief, as discussed in (Branchaud-Charron et al., 2019), where it was argued that these parameters had minimal influence. In contrast, our findings provide compelling evidence that both $K$ and $M$ play a critical role in shaping the spectral complexity assessment. Specifically, increasing $K$ leads to higher CSG values (in general), reflecting an increased perception of dataset complexity. It is likely that, smaller $K$ values tend to ignore large scale structural features (like connectivity patterns) within the data, thereby missing to capture fine-grained variations that capture complex interactions and class overlap. This results in a lower perceived complexity. The interaction between $K$ and $M$ also reveals key insights about their joint influence on complexity estimation. Specifically, for smaller $K$ values, $M$ plays a critical stabilizing role, as low $K$ is highly sensitive to sampling effects.

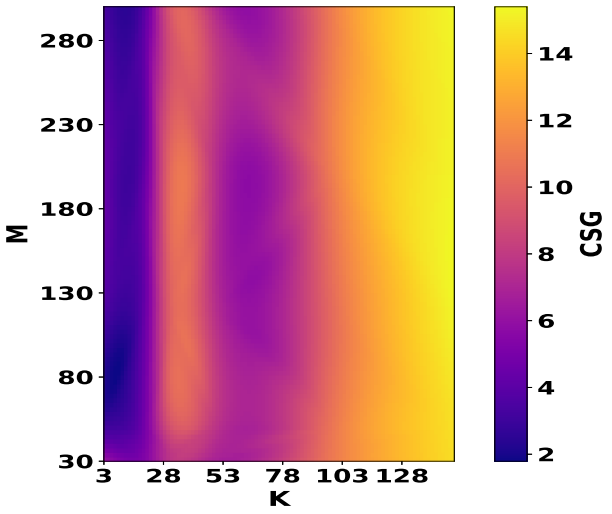


*Figure 2.* CSG as a function of $M$ and $K$ values.

Additionally, Figure 3 illustrates how the CSG is strongly influenced by the parameter $K$, with CSG values increasing consistently as $K$ increases across all datasets. This trend reveals that larger $K$-values capture broader structural patterns, leading to higher perceived complexity, while smaller $K$-values emphasize local structure and result in lower CSG. Furthermore, the variation in CSG across different datasets highlights the importance of tailoring $K$ to the specific structural and semantic characteristics of each KG.

Finally, Figure 4 plots CSG values for five standard KG benchmarks against the corresponding MRR values achieved by a suite of tail-prediction models. Contrary to Branchaud-Charron et al. (2019), we observe no meaningful correlation (mean Pearson coefficient $R = -0.644$) between CSG and model performance across all datasets and methods. In summary, contrary to the assertion that CSG can consistently forecast downstream performance, these findings cast doubt on its value as a model-independent separability measure for large-scale classification tasks specially in KG domain and underscore the necessity for more reliable metrics in KG evaluation.

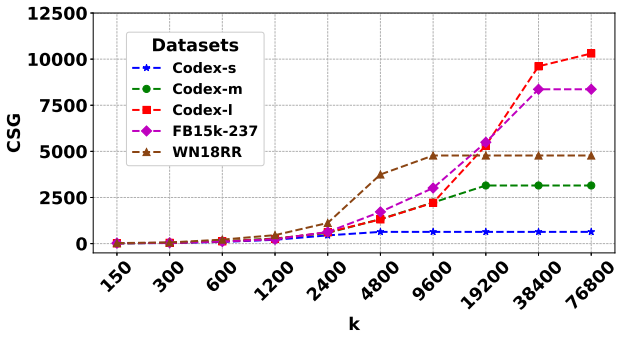


*Figure 3.* A plot of CSG as a function of $K$ values at $M = 100$.

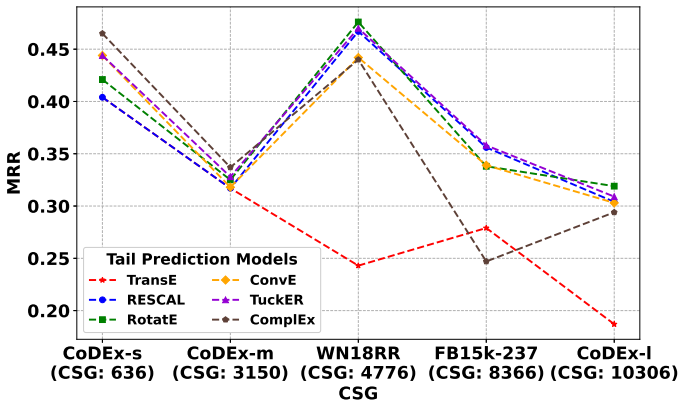


*Figure 4.* Relationship Between MRR from different tail-prediction models on five standard KG datasets and the corresponding CSG values.

## 3. Conclusion

CSG is significantly influenced by parameters, $K$ and $M$, challenging prior assumptions of their minimal impact on complexity assessments as well as the application of CSG as a reliable complexity metric for large multi-class datasets.

## References


Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270, 2004. doi: 10.1093/nar/gkh061. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proc. of ACM SIGMOD*, 2008.

Bourli, S. and Pitoura, E. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 6–10. IEEE, 2020.

Branchaud-Charron, F., Achkar, A., and Jodoin, P.-M. Spectral metric for dataset complexity assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3215–3224, 2019.

Gul, H., Naim, A. G., and Bhat, A. A. A contextualized bert model for knowledge graph completion. In *Muslims in Machine Learning (MusIML) Workshop, Co-located with Advances in neural information processing systems-24 (NeurIPS'24), https://arxiv.org/html/2412.11016v1*, volume 3, 2024.

Gul, H., Naim, A. G., and Bhat, A. A. Mucos: Efficient drug target discovery via multi context aware sampling in knowledge graphs. In *24th BioNLP workshop, Co-located with Association for Computational Linguistics (ACL-2025), https://arxiv.org/pdf/2503.08075*, 2025a.

Gul, H., Naim, A. G. H., and Bhat, A. A. Mucokgc: Multi-context-aware knowledge graph completion. In *29th edition of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), https://arxiv.org/html/2503.03091v2*, 2025b.

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F., and He, K. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024. doi: 10.1109/TPAMI.2024.3417451.

Miller, G. A. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11), 1995.

Safavi, T. and Koutra, D. Codex: A comprehensive knowledge graph completion benchmark. In *Proc. of EMNLP*, 2020.

Spillo, G., Bottalico, F., Musto, C., De Gemmis, M., Lops, P., and Semeraro, G. Evaluating content-based pretraining strategies for a knowledge-aware recommender system based on graph neural networks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 165–171, 2024.

Zeng, Z., Cheng, Q., Hu, X., Zhuang, Y., Liu, X., He, K., and Liu, Z. Kosel: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems*, 309:112837, 2025.

Zhang, Y., Sui, X., Pan, F., Yu, K., Li, K., Tian, S., Erdengasileng, A., Han, Q., Wang, W., Wang, J., et al. A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. *Nature Machine Intelligence*, pp. 1–13, 2025.