StratGraph: A Cost-Efficient Graph-RAG Framework via Structural Stratified Sampling

Chongchong Yang, Chaoqian Liu, Zhiqiang Yang, and Wenhao Zhu*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
{0124ycc, shengfang527, yangzq, whzhu}@shu.edu.cn (*Corresponding author)

***Abstract*—Graph-RAG enhances LLM reasoning but incurs prohibitive indexing costs. Existing cost-efficient methods, which rely on global centrality metrics like PageRank, introduce a systematic "Hub Bias." This bias favors information from dominant topics at the expense of discarding critical evidence from long-tail semantic clusters. To overcome this limitation, we propose StratGraph, a novel framework that prioritizes structural diversity over centrality through its core Structural Stratified Sampling (S^3) strategy. S^3 enforces proportional budget allocation across semantically partitioned communities, ensuring the retention of essential "bridging evidence." This evidence forms a robust knowledge skeleton within StratGraph's dual-granularity index. Extensive experiments on the HotpotQA and MuSiQue benchmarks demonstrate that StratGraph consistently outperforms centrality-based baselines under equivalent budgets. Notably, StratGraph demonstrates superior cost-efficiency, outperforming the strongest baseline even with a 20% reduced indexing budget. On the challenging MuSiQue benchmark, it achieves a 16.1% relative improvement in Exact Match, highlighting the critical role of mitigating Hub Bias for reliable multi-hop reasoning.**

***Index Terms*—Retrieval-Augmented Generation, Structural Stratified Sampling, Multi-hop Reasoning, Cost-Efficiency**

I. Introduction

Retrieval-Augmented Generation enhances the factual grounding of Large Language Models by retrieving evidence from external knowledge sources, thereby reducing hallucination. In contrast to traditional text-based RAG, which retrieves isolated chunks suitable for simple queries, Graph-RAG [1] is specifically designed for complex reasoning. It constructs a Text-Attributed Graph (TAG) from the corpus to explicitly model relational paths, enabling state-of-the-art performance on multi-hop question-answering tasks.

However, the construction of a full Knowledge Graph is computationally prohibitive, often necessitating LLM processing of every text chunk [2], [3]. Recent cost-efficient approaches (e.g., KET-RAG [4]) address this by pruning the graph based on global centrality metrics such as PageRank. We contend that this strategy introduces a systematic "Hub Bias." This issue arises from the inherently heavy-tailed structure of real-world knowledge graphs [5]. As a result, global centrality metrics disproportionately favor dominant topical hubs, systematically suppressing critical evidence from peripheral long-tail semantic regions.

This bias fundamentally undermines multi-hop QA, as the essential "bridging evidence" for connecting answers often lies within niche, low-centrality clusters. Centrality-based pruning, by discarding these peripheral nodes, breaks the reasoning pathways needed for complex questions, resulting in substantial recall loss on datasets like MuSiQue.

To address this limitation, we introduce StratGraph, a framework that explicitly shifts the indexing paradigm from global centrality ranking to structural diversity-aware selection. Its key innovation is S^3, which partitions the corpus into semantic communities and enforces a proportional budget allocation. By ensuring that core chunks are selected from all communities—especially the long-tail ones marginalized by centrality metrics—StratGraph builds a balanced, dual-granularity index. This approach effectively reconciles the need for high-level reasoning with the necessity of broad lexical coverage.

The experiments on HotpotQA and MuSiQue demonstrated that, under the same budget, StratGraph consistently outperformed the centrality-based baseline. Crucially, our method exhibits exceptional cost-efficiency: StratGraph surpasses the strongest baseline even when reducing the indexing budget by 20%. Notably, on MuSiQue, it achieved a relative improvement of 16.1% in exact matching compared to KET-RAG, empirically confirming the necessity of mitigating Hub Bias for deep reasoning. The implementation code and experimental data are publicly available in our GitHub repository.

The main contributions of this work are threefold:

- We identify a fundamental limitation—Hub Bias—in existing cost-efficient Graph-RAG methods, which we formally demonstrate leads to the systematic loss of long-tail semantic evidence.
- We propose **StratGraph** with the S^3 strategy, which reconciles the trade-off between indexing cost and semantic coverage by enforcing diversity-aware budget allocation.
- Experiments confirm that StratGraph stands out in cost-efficient Graph-RAG systems, consistently retrieving the critical boundary cases for multi-hop reasoning.

II. Related Work

In this section, we review the evolution of RAG systems, categorizing them into graph-enhanced architectures, cost-efficient indexing, and clustering-based strategies. We highlight the limitations of existing approaches—specifically the "Hub Bias"—that motivate our StratGraph framework.

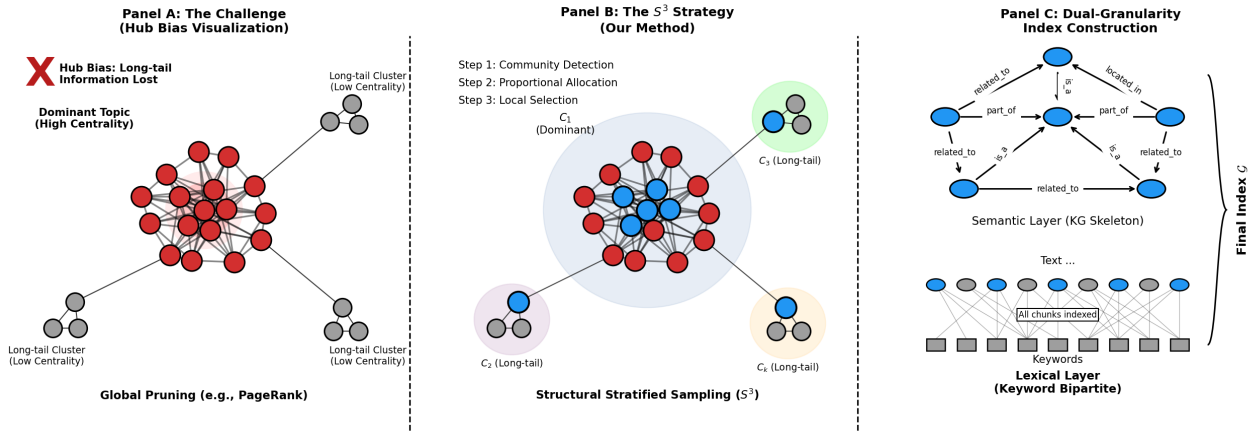

Fig. 1. Overview of StratGraph. (A) Centrality pruning induces Hub Bias. (B) S^3 uses stratified sampling to ensure diverse core selection. (C) A dual-granularity index supports reasoning and lexical coverage.

From Text-RAG to Graph-RAG. Conventional Text-RAG [6] retrieves independent text chunks based on semantic similarity. While effective for simple queries, this paradigm struggles with multi-hop reasoning, which requires synthesizing evidence across disparate documents. To bridge this gap, Graph-RAG structures the corpus into a Text-Attributed Graph (TAG). Early methods constructed implicit graphs (e.g., KNN graphs) based on similarity [7], which lack the precision to model explicit relationships. State-of-the-art systems, such as Microsoft's GraphRAG [1] and HippoRAG [8], employ LLMs to extract explicit relational triplets, enabling superior reasoning through structured retrieval. However, constructing a comprehensive knowledge graph necessitates processing every text chunk with an LLM, incurring prohibitive costs [9].

Efficient Indexing and the "Hub Bias." To mitigate these prohibitive indexing costs, recent research focuses on graph pruning and selective indexing [3], [10]. The core premise is to identify a small subset of "core" chunks for resource-intensive KG extraction. Cost-efficient methods, such as KET-RAG [4], predominantly select these core chunks based on Global Centrality metrics like PageRank. While effective at capturing dominant narrative hubs, we argue this introduces a systematic **Hub Bias**. Global metrics inherently amplify nodes within densely-connected clusters, systematically under-representing information from peripheral semantic clusters. This bias undermines multi-hop QA, as the crucial "bridging evidence" required to connect disparate facts is often located within these pruned, low-centrality regions.

Clustering and Community Detection in RAG. Clustering techniques serve two primary purposes in existing literature: hierarchical abstraction and retrieval optimization. In the embedding space, geometric clustering manages latent semantic topics (e.g., CoS [11] generates hierarchical summaries, while CRAG [12] performs group-level filtering). In the topological space, systems like Microsoft's GraphRAG [1] and HiRAG [13] apply community detection algorithms to construct multi-layered, hierarchical knowledge graphs. However, a shared characteristic is that clustering is used to augment an existing pipeline—either by constructing supplementary abstraction layers or compressing retrieved results. This introduces additional index maintenance and generative overhead. In contrast, StratGraph elevates clustering from an auxiliary tool to the primary mechanism for index-time resource allocation. By enforcing proportional budget constraints through semantic partitions, StratGraph fundamentally addresses the Hub Bias and ensures broad semantic coverage without relying on costly auxiliary layers.

III. Methodology

In this section, we propose StratGraph, a cost-efficient Graph-RAG framework designed to reconcile the trade-off between indexing cost and knowledge coverage. The core innovation of StratGraph is the S^3 strategy, which resolves the Hub Bias inherent in traditional centrality-based pruning methods. By ensuring equitable representation of diverse semantic communities, StratGraph constructs a high-quality Knowledge Graph (KG) Skeleton that supports complex reasoning with minimal token consumption.

A. Framework Overview

The StratGraph framework operates on a corpus of text chunks $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$. Our objective is to construct a Text-Attributed Graph index that supports multi-hop reasoning while constraining the computational cost of Large Language Model (LLM) extraction.

As illustrated in Fig. 1, the framework consists of three phases:

1) **Topology Initialization:** Modeling latent semantic relationships between chunks using a lightweight proxy graph.
2) **Stratified Skeleton Construction:** Identifying a budget-constrained set of "Core Chunks" ($\mathcal{V}_{\text{core}}$) via the S^3 algorithm to form the semantic backbone.
3) **Dual-Granularity Indexing:** Supplementing the KG skeleton with a fine-grained keyword-bipartite graph to ensure zero information loss.

B. Structural Stratified Sampling (S^3)

To address the limitations of global pruning, we introduce the S^3 algorithm to identify the core set $\mathcal{V}_{\text{core}}$. Unlike random sampling (which lacks focus) or global centrality (which lacks diversity), S^3 integrates **Community Detection** with **Local Centrality Ranking**.

1) Step 1: Semantic & Structural Partitioning: To capture the latent diversity of the corpus, we partition the text chunks into disjoint semantic communities $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. Recognizing that semantic manifolds can exhibit diverse structures, our framework supports two complementary partitioning paradigms:**Vector-Space Clustering (Geometric):** For corpora with distinct topical separation, we apply geometric clustering algorithms (e.g., K-Means) directly on the dense embeddings. Crucially, to avoid arbitrary hyperparameter selection, we employ the Silhouette Coefficient to dynamically determine the optimal number of clusters K. We iterate through a candidate range of K values and select the one that maximizes the Silhouette Score, ensuring that the resulting partitions exhibit high intra-cluster cohesion and inter-cluster separation. **Graph-Based Community Detection (Topological):** Alternatively, we model the chunks as a K-Nearest Neighbor (KNN) graph $G_{\text{proxy}} = (\mathcal{V}, \mathcal{E})$ and apply modularity-based algorithms (e.g., Leiden). This identifies structural communities based on information flow, where the community count is naturally determined by the modularity resolution.

Mathematically, regardless of the chosen paradigm, the corpus is decomposed into:

$$\mathcal{V} = \bigcup_{i=1}^{K} C_i, \quad C_i \cap C_j = \emptyset \quad \text{for } i \neq j. \tag{1}$$

2) Step 2: Proportional Budget Allocation: Let $B = \beta|\mathcal{V}|$ be the total budget. To maximize semantic coverage while preserving peripheral communities, we allocate a quota b_i to each community C_i proportional to its size, enforced by a lower-bound:

$$b_i = \max\left(1, \left\lfloor B \cdot \frac{|C_i|}{|\mathcal{V}|} \right\rfloor\right), \quad \text{where } \sum_{i=1}^{K} b_i \leq B. \tag{2}$$

3) Step 3: Intra-Community Ranking: Within each C_i, we perform a local sort to avoid redundancy. Based on global PageRank scores $\pi(v)$ from G_{proxy}, we select the top b_i representative nodes: $\mathcal{S}_i = \text{TopRank}(v \in C_i \mid \text{key} = \pi(v), \text{limit} = b_i)$. This identifies local hubs maintaining global relevance. The final core set is defined as $\mathcal{V}_{\text{core}} = \bigcup_{i=1}^{K} \mathcal{S}_i$ (summarized in Algorithm 1).

Algorithm 1 Structural Stratified Sampling (S^3)

Require: Corpus $\mathcal{T}$, Budget ratio $\beta \in (0, 1]$, Paradigm $P \in \{\text{Geom.}, \text{Topo.}\}$, Candidate K-range $[K_{\min}, K_{\max}]$ (if $P = \text{Geom.}$)
Ensure: Core chunk set $\mathcal{V}_{\text{core}} \subseteq \mathcal{T}$

GRAPH CONSTRUCTION
Build $G_{\text{proxy}} \leftarrow (\mathcal{T}, \mathcal{E})$ or embedding matrix $\mathbf{E}$
Compute $\pi(v) \leftarrow \text{PageRank}(v), \forall v \in \mathcal{T}$

SEMANTIC PARTITIONING
if $P = \text{Geom.}$ **then**
  $K^* \leftarrow \arg\max_k \text{Silhouette}(\mathbf{E}, k), \quad k \in [K_{\min}, K_{\max}]$
  $\mathcal{C} \leftarrow \text{K-Means}(\mathbf{E}, K^*)$
else
  $\mathcal{C} \leftarrow \text{Leiden}(G_{\text{proxy}})$
end if
$K \leftarrow |\mathcal{C}|, \quad \mathcal{C} \leftarrow \{C_1, \ldots, C_K\}$

STRATIFIED CORE SELECTION
$B \leftarrow \lceil \beta \cdot |\mathcal{T}| \rceil$ {Total budget}
$\mathcal{V}_{\text{core}} \leftarrow \emptyset$
for $i \leftarrow 1$ to K **do**
  $b_i \leftarrow \max\big(1, \lfloor B \cdot |C_i|/|\mathcal{T}| \rfloor\big)$
  $\mathcal{S}_i \leftarrow \underset{v \in C_i}{\text{Top-}} b_i\big(\pi(v)\big)$
  $\mathcal{V}_{\text{core}} \leftarrow \mathcal{V}_{\text{core}} \cup \mathcal{S}_i$
end for
return $\mathcal{V}_{\text{core}}$

C. Dual-Granularity Index Construction

StratGraph employs a hybrid indexing strategy designed to balance high-level reasoning with lexical precision. Specifically, we first construct a **KG Skeleton (Semantic Layer)** by employing an LLM to extract entities and relationships exclusively from the selected core set $\mathcal{V}_{\text{core}}$. Given the structural diversity of $\mathcal{V}_{\text{core}}$, the resulting knowledge graph G_{KG} effectively spans the distinct narrative threads of the corpus. To complement this high-level structure and capture fine-grained details, we simultaneously construct a **Keyword-Bipartite Graph (Lexical Layer)**, denoted as G_{bp}, which links all text chunks to their contained keywords. This layer serves as a lightweight fallback mechanism to ensure comprehensive coverage. The final unified index is formally defined as the union $\mathcal{G} = G_{\text{KG}} \cup G_{\text{bp}}$.

D. Hybrid Retrieval and Context Fusion

Upon constructing the dual-granularity index $\mathcal{G}$, StratGraph employs a dynamic hybrid retrieval mechanism to synthesize the final context for the LLM. To effectively balance high-level structural reasoning with fine-grained lexical grounding, we introduce a context allocation ratio $\theta \in [0, 1]$, which governs the distribution of the total context token budget L_{total} between the two indexing layers. The retrieval process operates via two parallel streams. First, the **Semantic Stream** performs a neighbor search on the Knowledge Graph G_{KG} to retrieve entity-relation paths representing reasoning chains, where the retrieved context C_{KG} is truncated to a budget of $L_{KG} = \lfloor \theta \cdot L_{total} \rfloor$. Simultaneously, the **Lexical Stream** queries the Keyword-Bipartite Graph G_{bp} to fetch specific text chunks containing precise keyword matches, constrained by the remaining budget $L_{kw} = L_{total} - L_{KG}$. The final input provided to the reasoning engine is the concatenation

of these two complementary contexts, formally denoted as $C_{final} = C_{KG} \oplus C_{kw}$. This mechanism allows StratGraph to flexibly shift its focus between structural connectivity and lexical precision based on the hyperparameter θ, the impact of which is analyzed in Sec. IV-C3.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the effectiveness of StratGraph. We first introduce the experimental setup, including datasets, metrics, and baselines. Subsequently, we report the comparative results on multi-hop QA benchmarks and perform ablation studies to verify the contribution of our core S^3 strategy.

A. Benchmarks and Baselines

Datasets and Metrics. We evaluate StratGraph on two multi-hop QA benchmarks: **HotpotQA** (distractor setting) [14], which tests noise filtering among distractor paragraphs, and **MuSiQue** [15], a stricter dataset requiring multi-step composition across disconnected semantic clusters—ideal for testing structural diversity. Performance is assessed via Retrieval Coverage(the proportion of QA instances where the retrieved context contains the ground-truth answer), Exact Match (EM), and F1 Score.

Baselines and Variants. We compare against **Microsoft GraphRAG** (Local Search) [1], representing the high-cost global summarization paradigm, and **KET-RAG** [4], our primary cost-efficient baseline that relies on global PageRank. Comparing with KET-RAG isolates the performance gain of replacing global centrality with structural diversity under identical budgets. Furthermore, we implement StratGraph using two partitioning paradigms to verify robustness: **StratGraph-Geo** operates in vector space (using K-Means, GMMs, or Agglomerative clustering), while **StratGraph-Topo** applies the Leiden algorithm on the proxy graph for topological community detection.

B. Implementation Details

For the embedding model, we utilize **text-embedding-v4**, a state-of-the-art embedding model from the Qwen family, to generate high-quality dense vectors for both text chunks and queries. For the generation component, we employ **Qwen3-32B** as the reasoning engine to generate final answers based on the retrieved context. All experiments are conducted with a unified indexing budget (β) across KET-RAG and StratGraph to ensure fair comparison. For detailed hyperparameter configurations and further implementation details, please refer to the open-source repository.

C. Experimental Results

Our comparative experiments on the MuSiQue and HotpotQA benchmarks yield the results summarized in Table I. These results consistently and decisively demonstrate the superiority of StratGraph over centrality-based pruning methods across all evaluation metrics.

1) Main Performance Comparison: The advantage is most pronounced on the complex, multi-hop MuSiQue dataset. Here, StratGraph-Topo (Leiden) achieves a remarkable 16.1% relative improvement in EM and a 9.1% increase in F1 score over the strongest baseline, KET-RAG. This performance gain is closely correlated with its higher Retrieval Coverage of 75.4% (vs. KET-RAG's 73.8%). This verifies that our S^3 strategy successfully retains the critical, long-tail bridging evidence necessary for rigorous multi-hop reasoning—evidence that PageRank-based methods systematically overlook.

TABLE I
MAIN RESULTS ON MULTI-HOP QA BENCHMARKS. WE COMPARE STRATGRAPH VARIANTS AGAINST STATE-OF-THE-ART BASELINES. **BOLD** INDICATES THE BEST PERFORMANCE AMONG COST-EFFICIENT METHODS. ALL STRATGRAPH VARIANTS OPERATE UNDER THE SAME BUDGET CONSTRAINT (β = 0.8) AS KET-RAG. (SG: STRATGRAPH, PR: PAGERANK, KM: K-MEANS, AGG: AGGLOMERATIVE, LEI: LEIDEN)

Method	MuSiQue			HotpotQA		
	Cov.	EM	F1	Cov.	EM	F1
MS GraphRAG	0.492	0.118	0.169	0.908	0.408	0.537
KET-RAG (PR)	0.738	0.174	0.254	0.954	0.426	0.555
SG-Geo (KM)	0.746	0.192	0.265	**0.968**	**0.446**	**0.579**
SG-Geo (Agg)	0.748	0.192	0.260	0.956	0.434	0.561
SG-Geo (GMM)	0.736	0.180	0.257	0.958	0.436	0.561
SG-Topo (Lei)	**0.754**	**0.202**	**0.277**	0.956	0.436	0.572

On HotpotQA, StratGraph continues to push the performance frontier despite strong baseline scores. StratGraph-Geo (K-Means) attains the highest scores across all metrics, improving EM by 4.7% and F1 by 4.3% over KET-RAG. Its Retrieval Coverage peaks at an exceptional 96.8%, substantially outperforming both Microsoft GraphRAG (90.8%) and KET-RAG (95.4%). These results confirm that StratGraph's benefits are twofold: it excels at identifying structural bridges in complex, topological graphs (as on MuSiQue), while also demonstrating exceptional robustness in filtering noise within dense, geometric embedding spaces (as on HotpotQA). Consequently, StratGraph delivers state-of-the-art performance that is both high and consistent, irrespective of underlying dataset characteristics.

2) Qualitative Analysis: Visualizing Structural Diversity: To qualitatively validate the enhanced structural diversity of our approach, we visualize the semantic distribution of embedded text chunks from the MuSiQue dataset using t-SNE. We compare the chunk selection patterns of our diversity-aware sampling strategy against the centrality-based PageRank baseline under a fixed budget of 20.0%.

As illustrated in Fig. 2, the two strategies exhibit starkly different resource allocation behaviors. The PageRank baseline (blue dots) concentrates on the high-density cores of dominant semantic clusters, leading to structural redundancy where multiple selected chunks encode similar information, while large portions of the semantic long tail remain unsampled.

In contrast, StratGraph (orange dots) disperses its sampling budget across the semantic space. By enforcing proportional allocation at the community level, our method probes periph-

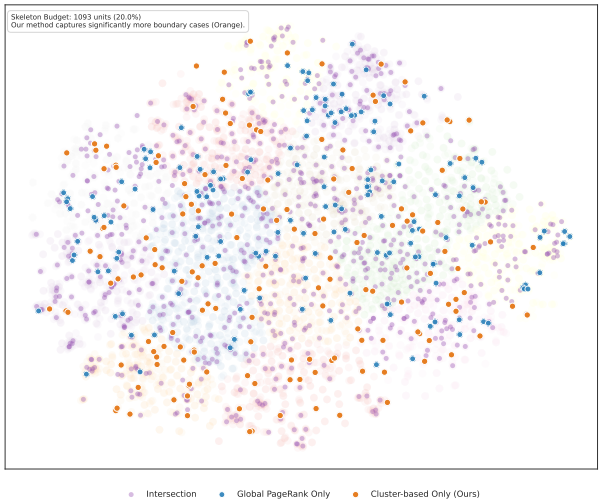

Fig. 2. Semantic space visualization of Knowledge Graph Skeleton. The semantic manifolds are represented by background clusters. Comparison between our cluster-based sampling (Orange) and the Global PageRank baseline (Blue) shows that our method achieves superior coverage in boundary regions while avoiding redundancy in high-density central clusters.

eral regions and captures critical bridging evidence at cluster boundaries—knowledge systematically deprioritized by global centrality metrics yet essential for forming coherent reasoning chains in multi-hop questions.

Furthermore, the overlapping selections (purple dots) confirm that StratGraph strikes a strategic balance: it retains access to essential global hubs while substantially expanding the semantic frontier of the resulting Knowledge Graph skeleton. This visualization provides direct empirical evidence that mitigating Hub Bias through structured diversity is a tangible and verifiable property of the constructed index.

3) Parameter Sensitivity Analysis: **Impact of Budget Ratio β.** Varying β from 0.2 to 1.0 (Fig. 3) reveals StratGraph's strong information retention under strict compression. Even at $\beta = 0.2$, StratGraph-Topo secures 94.4% Coverage and 0.526 F1, capturing critical structural bridges early. A cost-efficiency "sweet spot" emerges at $\beta = 0.4$, achieving 95% of full-index capability. Remarkably, StratGraph **reduces the indexing budget by 20%** against the baseline: at $\beta = 0.6$, both StratGraph-Geo (0.568 F1) and StratGraph-Topo (0.559 F1) surpass KET-RAG ($\beta = 0.8$, 0.555 F1). Beyond $\beta = 0.6$, diminishing returns confirm that S^3 effectively filters redundancy, enabling robust reasoning without a fully dense index.

Impact of Hyperparameter θ. Parameter θ controls context allocation between the KG Skeleton and Keyword index. Fig. 4 shows a clear "hybrid advantage" peaking at $\theta \in [0.2, 0.4]$, where balancing semantic reasoning and lexical matching allows StratGraph to consistently outperform KET-RAG. Furthermore, in high-θ regimes (> 0.6) with limited keyword evidence, StratGraph maintains superior stability. This confirms that our structural sampling constructs a significantly more robust and informative knowledge skeleton.

4) Ablation Study: Deconstructing the S^3 Strategy: To disentangle the respective contributions of structural partitioning and importance ranking, we conducted a multi-stage ablation study comparing StratGraph against two randomized variants: (1) Global Random, which samples chunks uniformly from the entire corpus under the same budget constraint, without any structural or semantic partitioning; and (2) Stratified Random, which employs our community partitioning but selects nodes randomly within each cluster.

TABLE II
STEP-WISE PERFORMANCE COMPARISON SHOWING THE BENEFITS OF STRATIFICATION (VS. GLOBAL RANDOM) AND LOCAL RANKING (VS. STRATIFIED RANDOM).

Method (Random)	MuSiQue			HotpotQA		
	Cov.	EM	F1	Cov.	EM	F1
Global Random	0.732	0.164	0.234	0.944	0.419	0.550
SG-Geo (KM)	0.736	0.168	0.239	0.962	0.426	0.562
SG-Geo (Agg)	0.732	0.166	0.235	0.955	0.424	0.559
SG-Geo (GMM)	0.729	0.160	0.227	0.958	0.412	0.554
SG-Topo (Lei)	0.736	0.178	0.250	0.964	0.428	0.559

Effectiveness of Structural Stratification. As shown in Table II, the Global Random baseline yields the lowest performance on MuSiQue (F1: 0.234, EM: 0.164). By merely introducing semantic partitioning, the Stratified Random variant (using Leiden) improves F1 to 0.250 and EM to 0.178. This performance gain suggests that structural constraints alone can partially mitigate Hub Bias by encouraging the retrieval of diverse, long-tail evidence that global uniform sampling often misses.

Effectiveness of Local Importance Ranking. However, diversity alone is insufficient. The full StratGraph model, which employs Local PageRank for intra-community selection, further boosts performance to an F1 of 0.277 and an EM of 0.202 on MuSiQue, representing a 10.8% relative improvement over the Stratified Random variant. These results empirically validate our two-step design philosophy: structural stratification ensures broad semantic coverage of the search space, while local importance ranking improves the depth and precision of the selected evidence.

V. CONCLUSION

We introduced StratGraph, a cost-efficient Graph-RAG framework that mitigates the "Hub Bias" inherent in centrality-based pruning. By shifting the indexing paradigm toward *structural diversity* via Structural Stratified Sampling (S^3), StratGraph ensures balanced coverage of long-tail semantic clusters, retaining critical bridging evidence for multi-hop reasoning. Evaluations on HotpotQA and MuSiQue demonstrate consistent superiority over baselines; notably, it achieves a 16.1% Exact Match improvement on MuSiQue and surpasses the strongest baseline even with a 20% reduced indexing budget. Ablations confirm the essential synergy between diversity-aware partitioning and intra-community ranking. Future work will explore incremental S^3 updates for dynamic corpora and adapt the framework to multi-modal knowledge graphs.

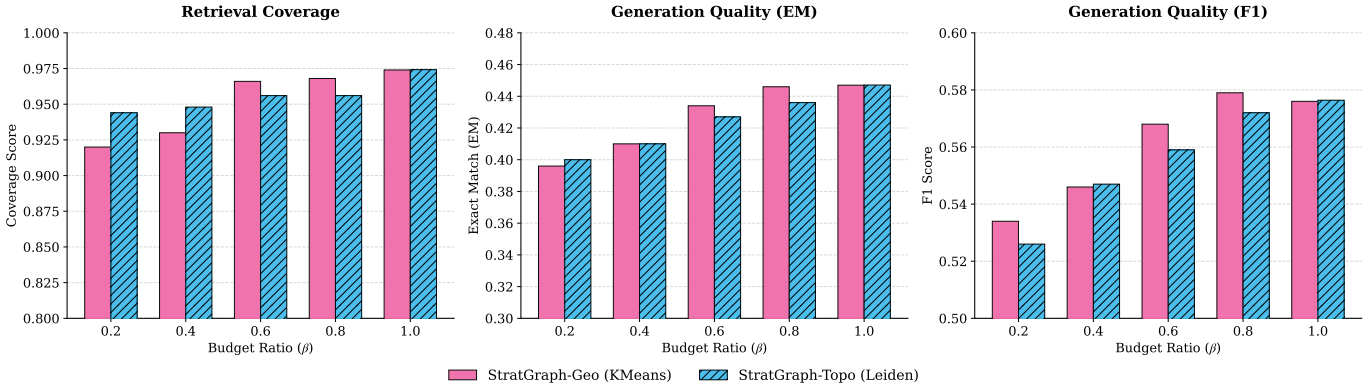

Fig. 3. Sensitivity to budget ratio β on HotpotQA. StratGraph achieves near-optimal cost-efficiency at $\beta = 0.4$, with diminishing returns at higher budgets.

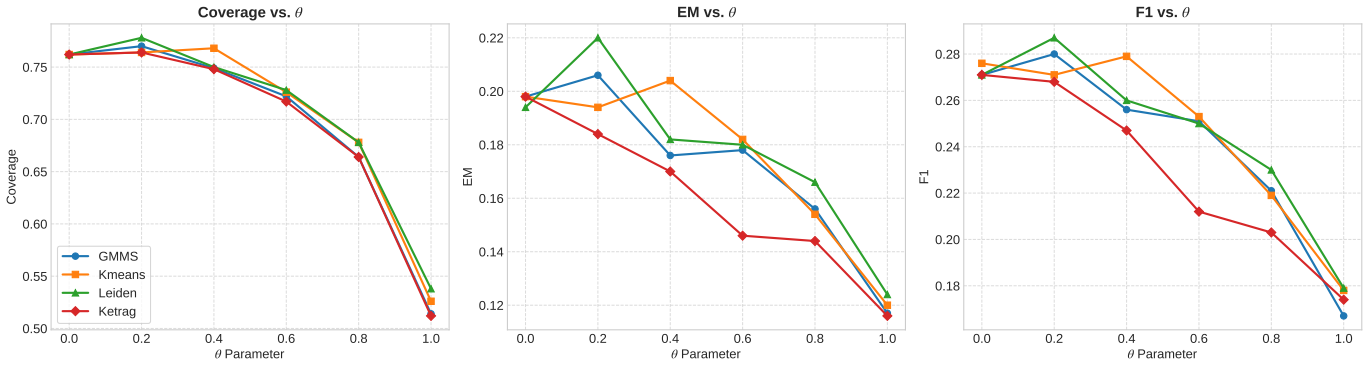

Fig. 4. Impact of Context Allocation Ratio θ on MuSiQue. Performance peaks at $\theta \in [0.2, 0.4]$, confirming the advantage of hybrid retrieval (balancing KG and Keywords). StratGraph variants consistently outperform the KET-RAG baseline across all effective ranges.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 12471500), and Shanghai Technical Service Computing Center of Science and Engineering, Shanghai University

REFERENCES

[1] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[2] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, H. Chen, Y. Xiao, C. Zhou, J. Dong *et al.*, "A survey of graph retrieval-augmented generation for customized large language models," *arXiv preprint arXiv:2501.13958*, 2025.

[3] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang, "Graph retrieval-augmented generation: A survey," *ACM Transactions on Information Systems*, 2024.

[4] Y. Huang, S. Zhang, and X. Xiao, "Ket-rag: A cost-efficient multi-granular indexing framework for graph-rag," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 1003–1012.

[5] H. Abdallah, B. Markhoff, and A. Soulet, "A complex network model for knowledge graphs' relationships," *Semantic Web*, vol. 16, no. 5, p. 22104968251361342, 2025.

[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[7] Z. Li, Q. Guo, J. Shao, L. Song, J. Bian, J. Zhang, and R. Wang, "Graph neural network enhanced retrieval for question answering of llms," *arXiv preprint arXiv:2406.06572*, 2024.

[8] B. Jimenez Gutierrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su, "Hipporag: Neurobiologically inspired long-term memory for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 532–59 569, 2024.

[9] S. Wang, Y. Fang, Y. Zhou, X. Liu, and Y. Ma, "Archrag: Attributed community-based hierarchical retrieval-augmented generation," *arXiv preprint arXiv:2502.09891*, 2025.

[10] C. Min, S. Bansal, J. Pan, A. Keshavarzi, R. Mathew, and A. V. Kannan, "Towards practical graphrag: Efficient knowledge graph construction and hybrid retrieval at scale," 2025. [Online]. Available: https://arxiv.org/abs/2507.03226

[11] C. Yang, W. Zhu, Z. Yang, C. Liu, and J. Hong, "Chain-of-summary: An efficient multi-clustering framework for hierarchical abstraction," in *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2025, pp. 673–682.

[12] S. Akesson and F. A. Santos, "Clustered retrieved augmented generation (crag)," *arXiv preprint arXiv:2406.00029*, 2024.

[13] H. Huang, Y. Huang, J. Yang, Z. Pan, Y. Chen, K. Ma, H. Chen, and J. Cheng, "Retrieval-augmented generation with hierarchical knowledge," *arXiv preprint arXiv:2503.10150*, 2025.

[14] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2369–2380.

[15] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Musique: Multihop questions via single-hop question composition," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 539–554, 2022.