Proof Sketch for "AGRAG: Advanced Graph-based Retrieval-Augmented Generation for LLMs"

Supplementary Material

Overview

This document serves as the supplementary proof sketch for the paper *"AGRAG: Advanced Graph-based Retrieval-Augmented Generation for LLMs"*. It provides the formal definition of the Minimum Cost Maximum Influence (MCMI) subgraph generation problem and the detailed proofs for the theorems regarding its complexity and the approximation guarantees of our proposed algorithm.

1 Problem Definition

We first assign node score $\mathcal{S}_\mathcal{V}$ and edge cost $\mathcal{C}_\mathcal{E}$ to Knowledge Graph (KG) $\mathcal{G}$. Based on the mapped triplet facts $\mathcal{F}$ and the KG structure, we calculate the Personalized PageRank (PPR) [1] score as our node influence score $\mathbf{s}$ for node set $\mathcal{V}$:

$$\mathbf{s} = \lim_{i\to\infty} \text{PPR}^i, \quad \text{PPR}^i = (1-d) \times \mathbf{p} + d \times \mathbf{P} \cdot \text{PPR}^{i-1}, \tag{1}$$

and the node score set $\mathcal{S}_\mathcal{V}$ is:

$$\mathcal{S}_\mathcal{V} = \{\mathbf{s}[v] \mid v \in \mathcal{V}\} \tag{2}$$

To align with [2], we set the damping factor d to 0.5, and converge the PPR calculation when it changes less than 10^{-7}; $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}|}$ denotes the personalization vector that depends on the mapped triples $\mathcal{F}$:

$$\mathbf{p}[u] = \begin{cases} 1, & \text{if } u \in \mathcal{F}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The $\mathbf{P} \in \mathbb{R}^{|\mathcal{V}\times\mathcal{V}|}$ in Equation (1) denotes the transition probability matrix, and can be obtained by:

$$\mathbf{P}[u][v] = \begin{cases} \frac{1}{|\mathcal{N}(u)|}, & \text{if } e_{(u,v)} \in \mathcal{E}, \\ \frac{1}{|\mathcal{V}|}, & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathcal{N}(u)$ denotes the 1-hop neighbor set of node u in the KG.

To ensure the PPR's faster convergence and more stable iterative computation [3], as well as normalizing the score distribution [4]. Following [2], we set the transition probability for non-edges to $\frac{1}{|\mathcal{V}|}$, and after calculate the PPR score, we multiply the PPR score of all passage nodes by a balance factor 0.05, to avoid over concentration on them. After that, we can obtain the node influence score s_v for each node $v \in \mathcal{V}$ as follow:

$$s_v = \mathbf{s}[v], \quad \forall v \in \mathcal{V}. \tag{5}$$

At last, we calculate the edge cost $\mathcal{C}_\mathcal{E}$ based on the similarity between query q's representation $\mathbf{q}$ and relation edge e's respective triplet facts f_e's representation $\mathbf{f}_e$:

$$\mathcal{C}_\mathcal{E} = \{c_{e_{(u,v)}} \mid e_{(u,v)} \in \mathcal{E}\}, \quad c_{e_{(u,v)}} = \frac{1 - \text{MS}(\mathbf{q}, \mathbf{f}_{e_{(u,v)}})}{2}. \tag{6}$$

Definition 1 (Minimum Cost Maximum Influence (MCMI) Subgraph Generation Problem)**.** *Given a node-weighted and edge-weighted, connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{S}_{\mathcal{V}}, \mathcal{C}_{\mathcal{E}})$ and a set of mapped triples $\mathcal{F}$, let $\mathcal{V}_{term}$ be the set of nodes in $\mathcal{F}$, called terminals. Let $r \in \mathcal{V}_{term}$ be a designated root terminal. The goal is to generate a subgraph $\mathcal{G}_{MCMI}(\mathcal{V}', \mathcal{E}', \mathcal{S}'_{\mathcal{V}}, \mathcal{C}'_{\mathcal{E}})$ that contains all terminals $\mathcal{V}_{term}$ such that every terminal $v \in \mathcal{V}_{term}$ is reachable from root terminal r, while minimizing the total cost-score ratio $r(\mathcal{G}_{MCMI})$:*

$$r(\mathcal{G}_{MCMI}) = \sum_{e_{(u,v)} \in \mathcal{E}'} \frac{c_{e_{(u,v)}}}{s_u + s_v}. \tag{7}$$

$s_v = \mathbf{s}[v]$ denotes the Personalized PageRank (PPR) score of node v, as defined in Equation (1)*, and $c_{e_{(u,v)}}$ represents the edge cost of $e_{(u,v)}$, as defined in Equation* (6)*.*

2 Theorems and Proofs

2.1 Hardness of MCMI Generation

Theorem 1. *The MCMI subgraph generation problem is NP-hard and its approximation ratio is unbounded.*

Proof. We prove NP-hardness by restricting MCMI to a special instance equivalent to the Steiner Tree problem [5]. Let the node score $s_v = 1$ for all $v \in \mathcal{V}$. The objective function simplifies to $r(\mathcal{G}_{MCMI}) = \sum_{e_{(u,v)} \in \mathcal{E}'} c_{e_{(u,v)}}$. Minimizing this sum to connect the terminal node set $\mathcal{V}_{term}$ is exactly the Steiner Tree problem. Since the Steiner Tree problem is NP-hard, MCMI is also NP-hard. Furthermore, regarding the approximation bound, the standard 2-approximation for Steiner Trees relies on the triangle inequality of edge weights. However, the effective weight in MCMI, defined as:

$$w_{e_{(u,v)}} = \frac{c_{e_{(u,v)}}}{s_u + s_v}. \tag{8}$$

depends on independent variables $c_{e_{(u,v)}} \in (0, 1)$ and $s_v \in (0, 1)$. We can construct an instance where $s_u, s_v \to 0$, causing $w_{e_{(u,v)}} \to \infty$ regardless of a fixed small $c_{e_{(u,v)}}$. This independence allows arbitrary violations of the triangle inequality, implying that the effective weights are not bounded by input costs, thus invalidating the metric-based constant approximation ratio.

□

2.2 Approximation Guarantee of MCST Step

Theorem 2. *The Minimum Cost Steiner Tree (MCST) generation step (Lines 1-10 in Algorithm 3 of the main paper) guarantees a 2-approximation.*

Proof. Suppose the cost for $\mathcal{G}_{MCST}(\mathcal{V}, \mathcal{E}, \mathcal{S}_V, \mathcal{C}_\epsilon)$ is $\gamma = \sum_{c_e \in \mathcal{C}_{\mathcal{E}}} c_e$, where $c_e \in \mathcal{C}_{\mathcal{E}}$ is the cost of edge e. Denote the optimal MCST as $\mathcal{G}^*(\mathcal{V}^*, \mathcal{E}^*, \mathcal{S}^*_{\mathcal{V}}, \mathcal{C}^*_{\mathcal{E}})$ with cost $\gamma^* = \sum_{c_e \in \mathcal{C}^*_{\mathcal{E}}} c_e$.

Following the triangle inequality [13], for any edge e in the optimal MCST's edge set $\mathcal{E}^*$, the cost of our approximated MCST is less than 2 times the optimal cost γ^*. That is:

$$\gamma \leq 2\gamma^* \tag{9}$$

Hence, the connectivity backbone cost is bounded by 2. This ensures that the fundamental reasoning paths connecting all query-mapped entities are retrieved with a bounded connectivity cost. □

2.3 Convergence of Greedy Expansion

Theorem 3. *The greedy expansion of the MCMI generation step (Lines 11-15 in Algorithm 3 of the main paper) strictly monotonically gains information and terminates within at most $|\Omega|$ iterations, where Ω denotes the candidate edge set of MCMI:*

$$\Omega = \{e_{(u,v)} \in \mathcal{E} \setminus \mathcal{E}_{MCST} \mid \frac{s_v}{c_{e_{(u,v)}}} > r(\mathcal{G}_{MCST})\} \tag{10}$$

Proof. Let $s_v/c_{e_{(u,v)}}$ be the marginal score-cost ratio of edge e. The greedy expansion of the MCMI generation step adds edge e if and only if its marginal score-cost ratio strictly exceeds the current graph's score-cost ratio $r(\mathcal{G}_{MCMI})$, i.e., $s_v/c_{e_{(u,v)}} > r(\mathcal{G}_{MCMI})$, where:

$$r(\mathcal{G}_{MCMI}) = \sum_{e_{(u,v)} \in \mathcal{E}_{MCMI}} \frac{s_u + s_v}{2 \times c_{e_{(u,v)}} \times |\mathcal{E}_{MCMI}|} \tag{11}$$

Since only the edges whose marginal score-cost ratio exceeds $r(\mathcal{G}_{MCMI})$ will be added, and all node scores s_u, s_v and edge costs $c_{e_{(u,v)}}$ are strictly greater than 0, $r(\mathcal{G}_{MCMI})$ monotonically increases.

After generating the MCST, let Ω be the set of edges whose marginal density ratio exceeds the graph density of the MCST and which do not belong to the MCST. The greedy expansion step takes at most $|\Omega|$ iterations to process each candidate edge. Since the edge set is finite, the algorithm is guaranteed to terminate. □

References

[1] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 517–526.

[2] B. J. Gutiérrez, Y. Shu, W. Qi, S. Zhou, and Y. Su, "From rag to memory: Non-parametric continual learning for large language models," *Forty-second International Conference on Machine Learning*, 2025.

[3] P. Boldi, M. Santini, and S. Vigna, "Pagerank as a function of the damping factor," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 557–566.

[4] R. Baeza-Yates, P. Boldi, and C. Castillo, "Generic damping functions for propagating importance in link-based ranking," *Internet Mathematics*, vol. 3, no. 4, pp. 445–478, 2006.

[5] E. N. Gilbert and H. O. Pollak, "Steiner minimal trees," *SIAM Journal on Applied Mathematics*, vol. 16, no. 1, pp. 1–29, 1968.