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.*

Proof. We first establish the equivalence between the MCMI problem and the classical Steiner Tree problem [5]. Let us define a transformed weight function $w : \mathcal{E} \to \mathbb{R}^+$ for the graph $\mathcal{G}$, where the weight of an edge $e_{(u,v)}$ is given by $w(e_{(u,v)}) = \frac{c_{e_{(u,v)}}}{s_u + s_v}$. Substituting this into the MCMI objective function (Equation (7)), we obtain $r(\mathcal{G}_{MCMI}) = \sum_{e \in \mathcal{E}'} w(e)$. This transformation establishes a bijection between the MCMI problem and the Steiner Tree problem on $\mathcal{G}$ with weights w. Since the connectivity constraints are topological, a subgraph $\mathcal{G}'$ connects the terminal set $\mathcal{V}_{term}$ in the MCMI instance if and only if it connects $\mathcal{V}_{term}$ in the Steiner Tree instance. Thus, a solution exists for one problem if and only if it exists for the other.

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 Effectiveness Guarantee

We further prove our MCMI generation algorithm's MCST generation step (Lines 1-10 in Algorithm 1) can still have a 2-approximation rate in Theorem 2. And its MCMI generation step (Lines 11-15 in Algorithm 1) strictly monotonically gains information and terminates with in at most $|\Omega|$ iterations in Theorem 3.

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

Proof. Proof. Let γ and γ^* denote the costs of the generated subgraph $\mathcal{G}_{\text{MCST}}$ and the optimal Steiner tree $\mathcal{G}^*$, respectively. The algorithm computes an MST on the auxiliary graph H, where edge costs $c^H_{e_{(u,v)}}$ of $e_{(u,v)}$ in auxiliary graph H correspond to shortest path costs between node u and node v in the original knowledge graph $\mathcal{G}$. Consider a depth-first traversal of $\mathcal{G}^*$, which traverses each edge in $\mathcal{E}^*$ twice, yielding a total cost of $2\gamma^*$. As in the auxiliary graph H, any path between two node u, v that pass

Algorithm 1: MCMI Subgraph Generation Algorithm

Input : $\mathcal{KG}(\mathcal{V}, \mathcal{E}, \mathcal{S}_{\mathcal{V}}, \mathcal{C}_{\mathcal{E}})$: The weighted KG;
$\mathcal{F}$: A set of query-mapped triplet facts in $\mathcal{KG}$.

Output : $\mathcal{G}_{\text{MCMI}}(\mathcal{V}_{\text{MCMI}}, \mathcal{E}_{\text{MCMI}})$: The MCMI subgraph.

1 Compute an approximated minimum cost Steiner tree (MCST) $\mathcal{G}$ of $\mathcal{KG}$;
2 Let $\mathcal{V}_{\text{term}} \leftarrow \mathcal{F} \cap \mathcal{V}_{\text{MCST}}$ be the terminal nodes in MCST;
3 Construct weighted auxiliary graph H: $\mathcal{V}_H \leftarrow \mathcal{V}_{\text{term}}$;
4 **for** *each pair of terminal nodes* $u, v \in \mathcal{V}_{term}$ **do**
5 Find the shortest path $P_{(u,v)}$ in $\mathcal{G}$ from u to v;
6 Let $c^H_{(u,v)} \leftarrow \sum_{e \in P_{(u,v)}} c_e$, where $c_e \in \mathcal{C}_{\mathcal{E}}$;
7 Compute an approximated MCST $\mathcal{G}'$ of H;
8 Let $\mathcal{V}_{\text{MCMI}} \leftarrow \emptyset$, $\mathcal{E}_{\text{MCMI}} \leftarrow \emptyset$;
9 **for** *each edge* $e_{(u,v)} \in \mathcal{G}'$ **do**
10 Add nodes and edges from shortest path $P_{(u,v)}$ in MCST to $\mathcal{V}_{\text{MCMI}}$ and $\mathcal{E}_{\text{MCMI}}$;
11 Calculate $r(\mathcal{G}_{\text{MCMI}})$ with Equation (7);
12 **while** *the node* v *in MCMI's 1-hop neighborhood with lowest* $c_{e_{(MCMI,v)}}/s_v$ *satisfies* $c_{e_{(MCMI,v)}}/s_v < r(\mathcal{G}_{MCMI})$ **do**
13 Recalculate $r(\mathcal{G}_{\text{MCMI}})$ with Equation (7);
14 $\mathcal{V}_{\text{MCMI}} \leftarrow \mathcal{V}_{\text{MCMI}} \cup \{v\}$;
15 $\mathcal{E}_{\text{MCMI}} \leftarrow \mathcal{E}_{\text{MCMI}} \cup \{e_{(\text{MCMI},v)}\}$;
16 **return** $\mathcal{G}_{MCMI}(\mathcal{V}_{MCMI}, \mathcal{E}_{MCMI})$;

another node o has cost $c^H_{e_{(u,o)}} + c^H_{e_{(o,v)}} \geq c^H_{e_{(u,v)}}$, the triangle inequality was satisfied. By the triangle inequality, shortcutting this traversal to visit only the terminals forms a Hamiltonian cycle in H have cost at most $2\gamma^*$ [5]. Since γ represents the minimum weight tree connecting these terminals in H, it is strictly bounded by the cost of this cycle. Thus, $\gamma \leq 2\gamma^*$, the MCST generation step (Lines 1-10 in Algorithm 1) guarantees a 2-approximation. □

□

Theorem 3. *The greedy expansion of the MCMI generation step (Lines 11-15 in Algorithm 1) strictly monotonically gains information and terminates within at most* $|\Omega|$ *iterations, where* Ω *denotes the candidate edge set of MCMI:*

$$\Omega = \left\{ e_{(u,v)} \in \mathcal{E} \setminus \mathcal{E}_{MCST} \mid \frac{c_{e_{(u,v)}}}{s_v} < r(\mathcal{G}_{MCST}) \right\} \tag{9}$$

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

$$r(\mathcal{G}_{\text{MCMI}}) = \sum_{e_{(u,v)} \in \mathcal{E}'} \frac{c_{e_{(u,v)}}}{s_u + s_v}. \tag{10}$$

Since only the edges whose marginal cost-score ratio smaller than $r(\mathcal{G}_{\text{MCMI}})$ will be added, and all of the node scores s_u, s_v and edge costs $c_{e_{(u,v)}}$ are greater than 0, $r(\mathcal{G}_{\text{MCMI}})$ monotonically decreases, which means the increase in influence score s_u, s_v and decrease in total edge cost $c_{e_{(u,v)}}$. After generates MCST, let Ω be the set of edges whose cost-score ratio smaller than the MCST's cost-score ratio by R-Equation (10) and does not belong to MCST. The greedy expansion of the MCMI generation step takes at most $|\Omega|$ iterations to process each candidate edge. □

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.