COREKG: Coreset-Guided Personalized Summarization of Knowledge Graphs

Anonymous Authors

A Theoretical Analysis

In the theoretical analysis, personalization is captured by fixing a single user and defining all variables with respect to that user's query workload Q. The sensitivity scores $s(t)$, sampling probabilities $p(t)$, total sensitivity S, and the coreset cost functions are all computed solely with respect to this user-specific workload, rather than a global query set. As a result, the approximation guarantees, variance bounds, and coreset size results are derived on a per-user basis and hold independently for each personalized summary. This ensures that the user-specific construction of COREKG is fully aligned with the theoretical analysis.

Let the graph $G = (V, E, T)$ be a knowledge graph, where V is the set of nodes, E is the set of edges, and $T \subseteq V \times E \times V$ is the set of triples. Let Q denote a user-specific query workload. For each query $q \in Q$, let $f_q(t) \in \{0, 1\}$ indicate whether a triple $t \in T$ is relevant to query q by calculating the sensitivity score. Using this, we define the total cost of the graph G concerning the query set Q as:

$$\text{cost}(G, Q) = \sum_{t \in T} \sum_{q \in Q} f_q(t) \tag{1}$$

This cost function measures the total number of associations between triples in T and queries in Q, which effectively counts the number of triples that are relevant to at least one query. A uniform weight $w(t) = 1$ is assigned to every triple in the full graph G.

To reduce computational or storage overhead, our goal is to approximate $\text{cost}(G, Q)$ using a compact, weighted subset of the original graph. We define this subset as a *coreset*, denoted $C = (V', E', T')$, where $V' \subseteq V$, $E' \subseteq E$, and $T' \subseteq T$ is a weighted subset of triples. Each triple $t \in T'$ is assigned a non-negative weight $w(t)$, representing its importance. The total cost of the coreset C with respect to the query set Q is defined as:

$$\text{cost}(C, Q) = \sum_{t \in T'} \sum_{q \in Q} w(t) \cdot f_q(t) \tag{2}$$

Definition 1 (Coreset)**.** Let $G = (V, E, T)$ be a knowledge graph and let $\varepsilon \in (0, 1)$ be an approximation parameter. A weighted subset $C = (V', E', T')$ where $V' \subseteq V$ and $E' \subseteq E$ and $T' \subseteq T$, is called an ε-*coreset* of G with respect to a query workload Q if, with probability at least $1 - \delta$, the following inequality holds:

$$|\text{cost}(G, Q) - \text{cost}(C, Q)| \leq \varepsilon \cdot \text{cost}(G, Q) \tag{3}$$

In other words, the coreset C provides a $(1 \pm \varepsilon)$ multiplicative approximation to the total cost of the full graph G, ensuring that query-relevant information is preserved up to a small relative error.

Instead of assigning a new function to model this, we interpret the existing cost formulation as inherently reflecting the weighted influence of each triple through its frequency and relevance across the query set. This perspective allows us to approximate $\text{cost}(G, Q)$ using a compact, weighted subset of triples selected via sensitivity-based sampling. The objective is to construct a coreset C such that.

$$\text{cost}(C, Q) \approx \text{cost}(G, Q),$$

Thereby enabling accurate and efficient summarization while preserving user-specific query semantics.

Now, we formally define the sensitivity score for every triple $t \in T$ that quantifies its importance with respect to the query workload Q.

Lemma 1 (Bound on Total Sensitivity)**.** *Let $G = (V, E, T)$ be a knowledge graph and Q a query workload. For each triple $t \in T$, we define its* sensitivity score *with respect to Q as*

$$s(t) = \sum_{q \in Q} \frac{f_q(t)}{\text{cost}(G, Q)},$$

where $f_q(t) \in 0, 1$ is an indicator of whether triple t is relevant to query q. Then, the total sensitivity across all triples is exactly the number of queries:

$$\sum_{t \in T} s(t) = |Q|.$$

Proof. Expanding the definition of $s(t)$:

$$\sum_{t \in T} s(t) = \sum_{t \in T} \sum_{q \in Q} \frac{f_q(t)}{\text{cost}(G, Q)}.$$

For a fixed query q, the inner sum counts $|T_q|$ contributions of $1/\text{cost}(G, Q)$, giving:

$$\sum_{t \in T} \sum_{q \in Q} \frac{f_q(t)}{\text{cost}(G, Q)} = \sum_{t \in T} \frac{|T_q|}{\text{cost}(G, Q)}.$$

Therefore:

$$\sum_{t\in T} s(t) = \sum_{t\in T} \frac{|T_q|}{\text{cost}(G,Q)}.$$

But by definition, $\sum_{q\in Q} |T_q| = \text{cost}(G,Q)$. Thus:

$$\sum_{t\in T} s(t) = \sum_{t\in T} \frac{\text{cost}(G,Q)}{\text{cost}(G,Q)} = |Q|.$$

□

Lemma 2 (Expectation and Variance of Coreset Cost)**.** *Let $G = (V, E, T)$ be a knowledge graph and Q a query workload. For each triple $t \in T$, let $p(t)$ be its sampling probability, X_t the number of times t is sampled in m independent rounds, and $w(t) = \frac{1}{m\cdot p(t)}$ its assigned weight. Define the cost contribution of t as*

$$|T_q| = \sum_{q\in Q} f_q(t),$$

where $f_q(t) \in \{0, 1\}$ denotes whether t is relevant to query q. Then, the following hold:

1. The expected cost of the coreset equals the cost of the full graph:

$$\mathbb{E}[\text{cost}(C,Q)] = \text{cost}(G,Q).$$

2. The variance of the coreset cost is bounded as:

$$\text{Var}[\text{cost}(C,Q)] \le \frac{S}{m} \cdot \text{cost}(G,Q)^2,$$

where $S = \sum_{t\in T} s(t)$ is the total sensitivity and m is the number of sampling rounds.

Proof. **Expectation.** The coreset is formed by sampling m triples with replacement, where each triple $t \in T$ is chosen with probability $p(t)$. Thus $X_t \sim \text{Binomial}(m, p(t))$. The coreset cost is

$$\text{cost}(C,Q) = \sum_{t\in T} X_t \cdot w(t) \cdot |T_q| = \sum_{t\in T} X_t \cdot \frac{1}{m\cdot p(t)} \cdot \sum_{q\in Q} f_q(t).$$

Taking expectation and applying linearity:

$$\mathbb{E}[\text{cost}(C,Q)] = \sum_{t\in T} \frac{1}{m\cdot p(t)} \cdot \sum_{q\in Q} f_q(t) \cdot \mathbb{E}[X_t].$$

we have $\mathbb{E}[X_t] = m \cdot p(t)$. Substituting:

$$\mathbb{E}[\text{cost}(C,Q)] = \sum_{t\in T} \sum_{q\in Q} f_q(t) = \text{cost}(G,Q).$$

Variance. Define per-round contribution $r_i = \sum_{t\in T} y_{ti} \cdot \frac{|T_q|}{m\cdot p(t)}$, where $y_{ti} \in \{0, 1\}$ indicates whether t was chosen in round i. Thus $\text{cost}(C,Q) = \sum_{i=1}^{m} r_i$. Expectation of r_i is

$$\mathbb{E}[r_i] = \frac{1}{m} \sum_{t\in T} |T_q|.$$

The variance is

$$\text{Var}[r_i] = \frac{1}{m^2} \sum_{t\in T} \frac{|T_q|^2}{p(t)} - \frac{\text{cost}(G,Q)^2}{m^2}.$$

Hence,

$$\text{Var}[\text{cost}(C,Q)] = m\cdot\text{Var}[r_i] = \frac{1}{m} \sum_{t\in T} \frac{|T_q|^2}{p(t)} - \frac{\text{cost}(G,Q)^2}{m}.$$

With sensitivity-based sampling $p(t) = \frac{s(t)}{S}$, we obtain

$$\text{Var}[\text{cost}(C,Q)] = \frac{S}{m} \sum_{t\in T} \frac{|T_q|^2}{s(t)} - \frac{\text{cost}(G,Q)^2}{m}.$$

Bounding $\frac{|T_q|^2}{s(t)}$ via sensitivity definition ensures

$$\text{Var}[\text{cost}(C,Q)] \le \frac{S}{m} \cdot \text{cost}(G,Q)^2.$$

Thus, the coreset estimator is unbiased (expectation matches the true cost), and its variance decreases inversely with m, controlled by total sensitivity S. □

Lemma 3 (Bound on Cost-Sensitivity Ratio)**.** *For any triple $t \in T$, the following inequality holds:*

$$\frac{|T_q|^2}{s(t)} \le \sum_{q:t\in T_q} |T_q|$$

Proof. We apply the Cauchy-Schwarz inequality to scalar sequences over the set of queries in which triple t appears. Define for each such query q:

$$a_q = 1, \quad b_q = \frac{1}{\sqrt{|T_q|}}$$

Then the Cauchy-Schwarz inequality implies:

$$\left(\sum_{q:t\in T_q} a_q b_q\right)^2 \le \left(\sum_{q:t\in T_q} a_q^2\right)\left(\sum_{q:t\in T_q} b_q^2\right)$$

Substituting the values of a_q and b_q, we get:

$$\left(\sum_{q:t\in T_q} \frac{1}{\sqrt{|T_q|}}\right)^2 \le \left(\sum_{q:t\in T_q} 1\right)\left(\sum_{q:t\in T_q} \frac{1}{|T_q|}\right)$$

$$\Rightarrow \quad |T_q|^2 \le \left(\sum_{q:t\in T_q} |T_q|\right) \cdot s(t)$$

Rearranging the terms:

$$\frac{|T_q|^2}{s(t)} \le \sum_{q:t\in T_q} |T_q|$$

□

Summing this inequality over all triples $t \in T$ gives

$$\sum_{t\in T} \frac{|T_q|^2}{s(t)} \le \sum_{t\in T} \sum_{q:t\in T_q} |T_q|.$$

We can now reverse the order of summation, since each term is non-negative:

$$\sum_{t\in T} \sum_{q:t\in T_q} |T_q| = \sum_{q\in Q} \sum_{t\in T_q} |T_q| = \sum_{q\in Q} |T_q|^2.$$

By definition of the total workload cost, $\text{cost}(G,Q) = \sum_{q\in Q}|T_q|$, and therefore

$$\sum_{q\in Q}|T_q|^2 \le \big(\sum_{q\in Q}|T_q|\big)^2$$

$$\Rightarrow \sum_{q\in Q}|T_q|^2 \le \text{cost}(G,Q)^2.$$

Combining these results, we obtain

$$\sum_{t\in T}\frac{|T_q|^2}{s(t)} \le \text{cost}(G,Q)^2.$$

Let $S > 1$

$$\begin{aligned}\text{Var}[\text{cost}(C,Q)] &= \frac{1}{m}\sum_{t\in T}\frac{|T_q|^2}{s(t)} - \frac{\text{cost}(G,Q)^2}{m}\\ &\le \frac{S}{m}\,\text{cost}(G,Q)^2 - \frac{\text{cost}(G,Q)^2}{m}\\ &\le \frac{S}{m}\,\text{cost}(G,Q)^2.\end{aligned}$$

Lemma 4 (Bound on the Deviation)**.** *Let $r_i = \frac{|T_q|}{mp(t)}$ be the cost contribution from a single sample in the coreset. Then:*

$$\begin{aligned}|r_i - \mathbb{E}[r_i]| &= \left|\frac{|T_q|}{mp(t)} - \frac{X}{m}\right|\\ &= \frac{1}{m}\cdot\left|\frac{|T_q|}{p(t)} - X\right|\\ &\le \frac{1}{m}\cdot\frac{|T_q|}{p(t)} \quad \textit{(since } X \ge 0\textit{)}\\ &= \frac{1}{m}\cdot\frac{|T_q|S}{s(t)} \quad \textit{(substitute } p(t) = s(t)/S\textit{)}\\ &= \frac{S}{m}\cdot\frac{|T_q|}{s(t)}\end{aligned}$$

Now, assuming worst case $\frac{|T_q|}{s(t)} \le X$, we define:

$$b = \frac{SX}{m}$$

Theorem 1 (Coreset Size)**.** *Let C be a coreset constructed using sensitivity-based sampling over query workload Q, and let $X = \text{cost}(G,Q)$ be the true cost on the full graph. Then, with probability at least $1-\delta$, the following holds:*

$$\Pr\left(\text{cost}(C,Q) > (1+\epsilon)\cdot\text{cost}(G,Q)\right) \le \delta$$

provided the number of samples m satisfies:

$$m \ge \frac{8S}{\epsilon^2}\cdot\log\left(\frac{1}{\delta}\right)$$

where $S = \sum_{t\in T}s(t)$ is the total sensitivity.

Proof. We express the total coreset cost as a sum of independent contributions:

$$\text{cost}(C,Q) = \sum_{i=1}^{m} r_i$$

where each $r_i = \frac{|T_q|}{mp(t)}$ is the contribution from a single sampled triple t. From earlier results, we know:

$$\mathbb{E}[r_i] = \frac{X}{m}, \qquad \text{Var}[\text{cost}(C,Q)] \le \frac{S}{m}\text{cost}(G,Q)^2$$

And from the deviation bound:

$$|r_i - \mathbb{E}[r_i]| \le \frac{SX}{m}$$

Applying Bernstein's inequality:

$$\Pr\left(\sum_{i=1}^{m} r_i > X + \epsilon X\right) \le \exp\left(\frac{-\epsilon^2\text{cost}(G,Q)^2}{2\cdot\text{Var}[\text{cost}(C,Q)] + \frac{2}{3}b\epsilon X}\right)$$

Substituting the bounds:

$$\text{Var}[\text{cost}(C,Q)] \le \frac{S}{m}\text{cost}(G,Q)^2, \qquad b \le \frac{SX}{m}$$

We obtain:

$$\begin{aligned}&\Pr\left(\text{cost}(C,Q) > (1+\epsilon)X\right)\\ &\le \exp\left(\frac{-\epsilon^2\,\text{cost}(G,Q)^2}{\frac{2S\,\text{cost}(G,Q)^2}{m} + \frac{2S\epsilon\,\text{cost}(G,Q)^2}{3m}}\right)\end{aligned}$$

Factoring out $\text{cost}(G,Q)^2$ and simplifying:

$$= \exp\left(\frac{-\epsilon^2 m}{2S + \frac{2}{3}\epsilon S}\right)$$

To ensure this probability is at most δ, we require:

$$\exp\left(\frac{-\epsilon^2 m}{2S + \frac{2}{3}\epsilon S}\right) \le \delta \Rightarrow m \ge \frac{2S + \frac{2}{3}\epsilon S}{\epsilon^2}\cdot\log\left(\frac{1}{\delta}\right)$$

Finally, noting that $2 + \frac{2}{3}\epsilon \le 8$ for $\epsilon \in (0,1)$, we conclude:

$$m \ge \frac{8S}{\epsilon^2}\cdot\log\left(\frac{1}{\delta}\right)$$

Substituting $S = |Q|$, we get:

$$m \ge \frac{8|Q|}{\epsilon^2}\cdot\log\left(\frac{1}{\delta}\right).$$

This follows from Bernstein's inequality: for any $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, choosing

$$m = O\left(\frac{S}{\varepsilon^2}\log\frac{1}{\delta}\right) = O\left(\frac{|Q|}{\varepsilon^2}\log\frac{1}{\delta}\right)$$

□

This follows from Bernstein's inequality since for any $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ choosing $m = O\big(\frac{S}{\varepsilon^2}\log\frac{1}{\delta}\big)$ guarantees with probability at least $1-\delta$ that the weighted coreset is a $(1\pm\varepsilon)$ approximation to the workload and hence the summary size depends on S (or equivalently $|Q|$) rather than $|T|$.

B Ablation Study

To evaluate the role of each component of COREKG, we perform an Ablation study in which we investigate the effect of each component on the overall performance while maintaining the same size summary. In this case, the first variant of the COREKG system is referred to as COREKG-Uniform. In this variant, we sample triples uniformly at random rather than using query-sensitive importance sampling to help identify the most relevant triples to the user-based query workloads. The analysis of this first variant demonstrates that COREKG-Uniform has significantly less coverage, which indicates that uniform random sampling is ineffective for identifying relevant triples. This is indicative of the need to use query-based sensitivity when preserving meaningful query structure.

The next variant we analysed is COREKG-Global. We were able to calculate the sensitivity across the entire query workload rather than on a per-user or per-workstream basis. COREKG-Global performed worse than the full COREKG model, demonstrating that to provide personalized query handling, we must include user-specific query behaviours as a fundamental component of sensitivity conditioning. In addition, both COREKG-Unweighted and COREKG used sensitivity sampling. However, COREKG-Unweighted treated each sample from the coreset equally and assigned each sample an equal weight. This means that it could not adequately handle sampling and thus produce an overall very low coverage. It has been demonstrated through the ablation study that Weighted Coreset Sampling is effective in reducing sampling bias and preserving query semantics across workloads.

Based on our ablation analysis, we conclude that these elements of COREKG are all important and complementary to COREKG's success.