Supplementary Materials for Query-based Graph Data Dynamic Pricing for Revenue Maximization

A The Detailed Proof

Theorem 1. *The pricing function $P(Q) = \sum_{v \in V_{G(Q)}} p(v) + \sum_{e \in E_{G(Q)}} p(e)$ is arbitrage-free.*

Proof. To demonstrate that the pricing function $P(Q)$ is arbitrage-free, we must show that a buyer cannot obtain the data corresponding to a query Q at a lower total cost by decomposing it into a set of smaller sub-queries. Formally, we will prove that the sum of the prices of the sub-queries is greater than or equal to the price of the original query.

Let a query Q be decomposed into a set of k sub-queries, $\{Q_1, Q_2, \dots, Q_k\}$. The sub-queries are not necessarily disjoint, meaning they may have overlapping vertices and edges. The union of the vertices and edges of these sub-queries reconstructs the original query Q:

1. The vertex set of Q is the union of the vertex sets of the sub-queries:
$$V(Q) = \bigcup_{i=1}^{k} V(Q_i)$$
2. The edge set of Q is the union of the edge sets of the sub-queries:
$$E(Q) = \bigcup_{i=1}^{k} E(Q_i)$$

The price of a direct query Q is defined as the sum of the base prices $p(\cdot)$ of all its unique constituent vertices and edges:

$$P(Q) = \sum_{v \in V(Q)} p(v) + \sum_{e \in E(Q)} p(e) \tag{Eq. A}$$

The total cost of the decomposition strategy is the sum of the prices of all individual sub-queries:

$$\sum_{i=1}^{k} P(Q_i) = \sum_{i=1}^{k} \left[\sum_{v \in V(Q_i)} p(v) + \sum_{e \in E(Q_i)} p(e) \right]$$

By rearranging the order of summation, this total cost can be expressed as:

$$\sum_{i=1}^{k} P(Q_i) = \left(\sum_{i=1}^{k} \sum_{v \in V(Q_i)} p(v) \right) + \left(\sum_{i=1}^{k} \sum_{e \in E(Q_i)} p(e) \right) \tag{Eq. B}$$

Let us analyze the vertex component of the total cost. Consider an arbitrary vertex $v \in V(Q)$. By the definition of set union, v must belong to at least one sub-query's vertex set $V(Q_i)$. Let N_v be the number of sub-queries that contain the vertex v. Due to potential overlaps, $N_v \geq 1$. In the summation $\sum_{i=1}^{k} \sum_{v \in V(Q_i)} p(v)$, the price $p(v)$ is counted N_v times. Therefore, the vertex component of Eq. (Eq. B) can be rewritten as:

$$\sum_{v \in V(Q)} N_v \cdot p(v)$$

Assuming base prices are non-negative ($p(v) \geq 0$), and since $N_v \geq 1$, it follows that:

$$\sum_{v \in V(Q)} N_v \cdot p(v) \geq \sum_{v \in V(Q)} 1 \cdot p(v)$$

Thus,

$$\sum_{i=1}^{k} \sum_{v \in V(Q_i)} p(v) \geq \sum_{v \in V(Q)} p(v) \tag{Eq. C}$$

Similarly, for an arbitrary edge $e \in E(Q)$, let N_e be the number of sub-queries containing it. We have $N_e \geq 1$. The edge component of the total cost in Eq. (Eq. B) is $\sum_{e \in E(Q)} N_e \cdot p(e)$. It follows that:

$$\sum_{i=1}^{k} \sum_{e \in E(Q_i)} p(e) \geq \sum_{e \in E(Q)} p(e) \tag{Eq. D}$$

By combining the inequalities (Eq. C) and (Eq. D), we can compare the total cost of the decomposed query (Eq. (Eq. B)) with the cost of the direct query (Eq. (Eq. A)):

$$\begin{aligned}\sum_{i=1}^{k} P(Q_i) &= \left(\sum_{i=1}^{k} \sum_{v \in V(Q_i)} p(v) \right) + \left(\sum_{i=1}^{k} \sum_{e \in E(Q_i)} p(e) \right) \\ &\geq \left(\sum_{v \in V(Q)} p(v) \right) + \left(\sum_{e \in E(Q)} p(e) \right) = P(Q)\end{aligned}$$

This leads to the final inequality:

$$\sum_{i=1}^{k} P(Q_i) \geq P(Q)$$

The inequality demonstrates that the total cost of purchasing a set of decomposed sub-queries is always greater than or equal to the cost of purchasing the single, encompassing query directly. Equality holds only in the specific case where all sub-queries are perfectly disjoint ($N_v = 1$ and $N_e = 1$ for all vertices and edges). In any case with overlap, the cost of decomposition will be strictly higher.

Since there is no financial incentive to decompose a query, the pricing function $P(Q)$ prevents this form of arbitrage and is therefore arbitrage-free. $\square$