Online supplemental materials accompanying the paper

Suppose a set of multiple top-k queries are $Q = \{q_1, q_2, ..., q_x\}$ with the corresponding sets $F = \{f_1, f_2, ..., f_x\}$ and $K = \{k_1, k_2, ..., k_x\}$. The largest element in F is f_{max}, and the largest element in K is k_{max}. Assume that the groups after combination are $\mathcal{G} = \{G_1, ..., G_g\}$. We change the way a little to show an execution plan. Suppose there is a valid execution plan $EP = \{(f'_1, k'_1), (f'_2, k'_2), ..., (f'_u, k'_u)\}$ where $f'_i = 2^{i-1} * f'_1$, $f'_u \leq f_{max}$, and $k'_u = k_{max}$. Moreover, we have $k'_i \leq k'_j$ where $i, j \leq u$ and $i < j$. That is, there are some queries executed according to frequency f'_i with the value of k'_i where $i \leq u$. Given this execution plan EP, if we want to execute G_i effectively, the time when G_i is first executed is $f^1_i = 2^{j-1} * f'_1$ where $2^{j-1} * f'_1 \leq f_i < 2^j * f'_1$, and $k^{max}_i < k'_j$. Let the cost-per-unit time with EP be m', and let m^*_g be the cost-per-unit time of the DP approach. Obviously, $m^*_g \leq m'$. Let m be the cost-per-unit time of no sharing approach.

First, we give an upper bound of m' in the following lemma. Then, we prove Theorem 7.1.

Lemma 11.1: $m' \leq 2 * k_{max}/f'_1$

Proof: For each element (f'_i, k'_i) in EP, the cost of executing the query with the k'_i value is k'_i. In a cycle of f'_u, the total cost of executing this query is $(f'_u/f'_i) * k'_i$. For all the elements in EP, the total cost is $\Sigma^u_{i=1}(f'_u/f'_i) * k'_i$. Thus, we have

$$m' = \Sigma^u_{i=1}(f'_u/f'_i) * k'_i/f'_u$$

As $f'_i = 2^{i-1} * f'_1$, we have

$$m' = \Sigma^u_{i=1} k'_i/(2^{i-1} * f'_1)$$

Since $k'_i \leq k_{max}$ and $\Sigma^u_{i=1} 1/2^{i-1} \leq 2$, we get

$$m' \leq 2 * k_{max}/f'_1$$

□

Proof of Theorem 4.1

Proof:

Obviously,

$$E(m/m^*_g|x, F, K) \geq E(m/m'|x, F, K) \tag{13}$$

According to conditional probability theory, we have

$$\begin{aligned} &E(m/m'|x, F, K) \\ &= \Sigma p(f'_1, k_{f'_1}|x, F, K) * E(m/m'|f'_1, k_{f'_1}, x, F, K) \\ &= \Sigma p(f'_1, k_{f'_1}|x, F, K) * \Sigma p(k'_1, k'_2, ..., k'_u|f'_1, k_{f'_1}, x, F, K) \\ &\quad * E(m/m'|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \end{aligned} \tag{14}$$

Replacing m' with the inequality in Lemma 11.1, we have

$$\begin{aligned} &E(m/m'|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \geq \\ &\quad E(m|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K)/(2 * k_{max}/f'_1) \end{aligned} \tag{15}$$

As each query in Q is independent, inequality (15) can be reduced to

$$\begin{aligned} &E(m/m'|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \geq x* \\ &\quad E(m_0|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K)/(2 * k_{max}/f'_1) \end{aligned} \tag{16}$$

where m_0 is the cost-per-unit time of one query executed with no sharing according to its frequency upper bound.

Let $A_i = \{(f, k)|2^{i-1} * f'_1 \leq f < 2^i * f'_1, 1 \leq k \leq k'_i\}$ where (f, k) is a point in the area A_i, $A_i \cap A_j = \varnothing$ and $1 \leq i \leq u-1$. Suppose $A_u = \{(f, k)|2^{u-1} * f'_1 \leq f < f_{max}, 1 \leq k \leq k'_u\}$. (f, k) can be considered as one query. A_i is shown in Figure 18. Suppose $A = \cup^u_{i=1} A_i$ and $|A|$ is the area of A which is the dashed area in Figure 18.

Given a query (f, k) in the area A, the cost-per-unit time is k/f. Suppose the probability of (f, k) existing in A is $p(f, k)$, we know that

$$\begin{aligned} &E(m_0|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \\ &\geq \Sigma_{(f,k)\in A} p(f, k) * k/f \end{aligned} \tag{17}$$

Obviously, $p(f, k) = 1/|A|$. Thus, inequality (17) can be reduced to

$$= 1/|A| * \Sigma_{(f,k)\in A} k/f \tag{18}$$

Since $A > \cup^{u-1}_{i=1} A_i$, we can reduce inequality (18) with inequality (19)

$$\geq 1/|A| * \Sigma^{u-1}_{i=1} \Sigma_{(f,k)\in A_i} k/f \tag{19}$$

As $2^{i-1} * f'_1 \leq f < 2^i * f'_1$ and $1 \leq k \leq k'_i$, inequality (19) can be reduced to

$$\begin{aligned} &\geq (1/|A|) * \Sigma^{u-1}_{i=1} \Sigma^{2^i * f'_1 - 1}_{f=2^{i-1} * f'_1} \Sigma^{k'_i}_{k=1} k/f \\ &\geq (1/|A|) * \Sigma^{u-1}_{i=1} ((k'_i)^2/2) * \Sigma^{2^i * f'_1 - 1}_{f=2^{i-1} * f'_1} 1/f \end{aligned} \tag{20}$$

According to the calculus theory, we have $\Sigma^{2^i * f'_1 - 1}_{f=2^{i-1} * f'_1} 1/f \geq ln\ (2^i * f'_1/(2^{i-1} * f'_1)$. Thus, inequality (20) can be reduced to inequality (21)

$$\begin{aligned} &\geq (1/|A|) * \Sigma^{u-1}_{i=1} ((k'_i)^2/2) * ln\ (2^i * f'_1/(2^{i-1} * f'_1)) \\ &= \Sigma^{u-1}_{i=1} (k'_i)^2/(2 * log_2\ e * |A|) \end{aligned} \tag{21}$$

According to the assumption, $k'_i \geq k'_1$ and $|A| \leq k'_u * f_{max}$. Thus, equality (21) can be reduced to

$$\geq (u-1) * (k'_1)^2/(2 * log_2\ e * k'_u * f_{max}) \tag{22}$$

Suppose the query with the frequency upper bound f'_1 has a $k_{f'_1}$ value. To satisfy this frequency upper bound, we know that $k'_1 \geq k_{f'_1}$. Thus, we can reduce inequality (22) to inequality (23).

$$\begin{aligned} &E(m_0|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \\ &\geq (u-1) * (k_{f'_1})^2/(2 * log_2\ e * k'_u * f_{max}) \end{aligned} \tag{23}$$

As $f'_u = 2^{u-1} * f'_1$ and $2^{u-1} * f'_1 \leq f_{max} < 2^u * f'_1$,

$$u \geq log_2\ f_{max}/f'_1 \tag{24}$$

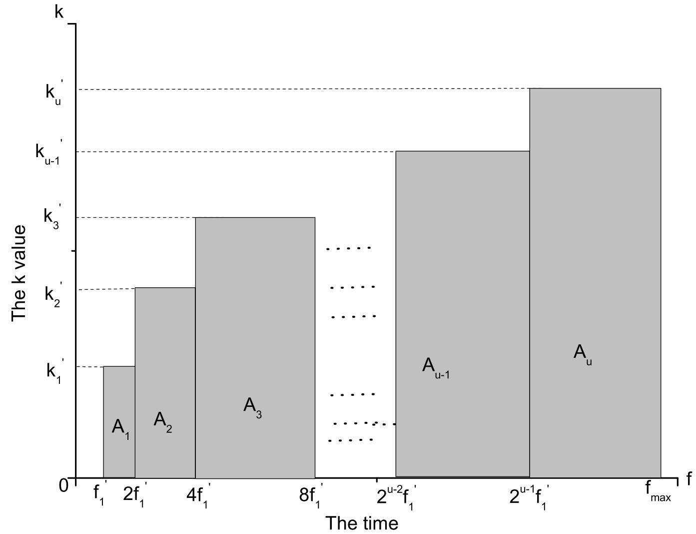


Fig. 18. The area of A_i and A where $1 \leq i \leq u$

Based on inequalities (23) and (24), and $k'_u = k_{max}$,

$$\begin{aligned} & E(m_0|k'_1, k'_2, ..., k'_u, f'_1, k_{f'_1}, x, F, K) \\ \geq \quad & (log_2\ f_{max}/f'_1 - 1) * ((k_{f'_1})^2)/(2 * log_2\ e * k'_u * f_{max}) \\ = \quad & (log_2\ f_{max}/f'_1 - 1) * ((k_{f'_1})^2) \\ & /(2 * log_2\ e * k_{max} * f_{max}) \end{aligned} \tag{25}$$

According to inequalities (13), (14), (16), and (25), we get

$$\begin{aligned} & E(m/m^*_g|x, F, K) \\ \geq \quad & \Sigma p(f'_1, k_{f'_1}|x, F, K) * \Sigma p(k'_1, k'_2, ..., k'_u|f'_1, k_{f'_1}, x, F, K) * \\ & x * (log_2\ f_{max}/f'_1 - 1) * (k_{f'_1})^2 \\ & /(2 * log_2\ e * k_{max} * f_{max}) * (2 * k_{max}/f'_1) \end{aligned} \tag{26}$$

For the above formula has nothing to do with $k'_1, k'_2, ..., k'_u$, we have

$$\Sigma p(k'_1, k'_2, ..., k'_u|f'_1, k_{f'_1}, x, F, K) = 1.$$

Thus, inequality (26) can be reduced to

$$\begin{aligned} = \quad & \Sigma p(f'_1, k_{f'_1}|x, F, K) * x * (log_2\ f_{max}/f'_1 - 1) * (k_{f'_1})^2 \\ & /(4 * log_2\ e * k^2_{max} * f_{max}/f'_1) \end{aligned} \tag{27}$$

As f'_1 and $k_{f'_1}$ are independent, we can reduce inequality (27) to equality (28).

$$\begin{aligned} = \quad & \Sigma_{f'_1=1}^{f_{max}} p(f'_1|x, F, K) * \Sigma_{k_{f'_1}=1}^{k_{max}} p(k_{f'_1}|x, F, K) * \\ & x * (log_2\ f_{max}/f'_1 - 1) * (k_{f'_1})^2 \\ & /(4 * log_2\ e * k^2_{max} * f_{max}/f'_1) \end{aligned} \tag{28}$$

Since $p(k_{f'_1}|x, F, K) = 1/k_{max}$, equality (28) can be reduced to

$$\begin{aligned} = \quad & \Sigma_{f'_1=1}^{f_{max}} p(f'_1|x, F, K) * x * (log_2\ f_{max}/f'_1 - 1) * \\ & \Sigma_{k_{f'_1}=1}^{k_{max}} (k_{f'_1})^2/(4 * log_2\ e * *k^3_{max} * f_{max}/f'_1) \\ \geq \quad & \Sigma_{f'_1=1}^{f_{max}} p(f'_1|x, F, K) * x * (log_2\ f_{max}/f'_1 - 1) \\ & /(12 * log_2\ e * f_{max}/f'_1) \end{aligned} \tag{29}$$

We shrink the range of f'_1. We handle it in two cases. When $f_{max}/2x \geq 1$, we reduce inequality (29) to

$$\begin{aligned} \geq \quad & \Sigma_{f'_1=f_{max}/2x}^{f_{max}/x} p(f'_1|x, F, K) * x * (log_2\ f_{max}/(f_{max}/x) - 1) \\ & /(12 * log_2\ e * f_{max}/(f_{max}/2x)) \\ = \quad & (p(f'_1 \geq f_{max}/2x) - p(f'_1 \geq f_{max}/x)) * (log_2\ x - 1) \\ & /(24 * log_2\ e) \\ = \quad & ((1 - 1/2x)^x - (1 - 1/x)^x) * (log_2\ x - 1)/(24 * log_2\ e) \\ = \quad & ((1/e)^{1/2} - 1/e) * (log_2\ x - 1)/(24 * log_2\ e) \\ = \quad & \Omega(log_2\ x) \end{aligned}$$

When $f_{max}/2x < 1$, as $f'_1 \geq 1$, we have $f_{max}/f'_1 < 2x$. Thus,

$$\begin{aligned} & (log_2\ f_{max}/f'_1 - 1)/(12 * log_2\ e * f_{max}/f'_1) \\ > \quad & (log_2 2x - 1)/(12 * log_2\ e * 2x) \end{aligned}$$

Then, inequality (29) can be reduced to

$$\begin{aligned} \geq \quad & \Sigma_{f'_1=1}^{f_{max}/x} p(f'_1|x, F, K) * x * (log_2\ 2x - 1) \\ & /(12 * log_2\ e * 2x) \\ = \quad & (p(f'_1 \geq f_{max}/2x) - p(f'_1 \geq f_{max}/x)) * (log_2\ 2x - 1) \\ & /(24 * log_2\ e) \\ = \quad & ((1 - 1/2x)^x - (1 - 1/x)^x) * (log_2\ 2x - 1) \\ & /(24 * log_2\ e) \\ = \quad & ((1/e)^{1/2} - 1/e) * (log_2\ 2x - 1)/(24 * log_2\ e) \\ = \quad & \Omega(log_2\ x) \end{aligned}$$

□

The analysis shows that the cost-per-unit time of no sharing is $\Omega(log_2\ x)$ times of that of DP in average, where x is the original number of queries in the system.

Proof of Lemma 5.1

Proof: Let t_{i-1}^1 be the first time to execute G_{i-1} in the execution with m_{i-1} where $t_{i-1}^1 \leq f_{i-1}$. As $f_{i-1} < f_i$ and $t_{i-1}^1 \leq f_{i-1}$, we have $t_{i-1}^1 < f_i$. Then, $f_i/t_{i-1}^1 \geq 1$ (Note that "/" in the formula is integer division operator). For $t_i^1 = (f_i/t_{i-1}^1)*t_{i-1}^1$, thus, $t_i^1 \geq t_{i-1}^1$. We know that $f_i-(f_i/t_{i-1}^1)*t_{i-1}^1 \leq t_{i-1}^1$, that is $f_i - t_i^1 \leq t_{i-1}^1$. We have $t_i^1 \geq f_i - t_{i-1}^1$

As $t_i^1 \geq t_{i-1}^1$ and $t_i^1 \geq f_i - t_{i-1}^1$, we have $2t_i^1 \geq t_{i-1}^1 + f_i - t_{i-1}^1 = f_i$. Thus, $t_i^1 \geq f_i/2$ □

Proof of Theorem 5.1

Proof: According to Equation (11), we know that $GAcost(t_i^1) = GAcost(t_i^1 - 1) + k_i^{max}$. Thus, we have

$$m_i = (GAcost(t_i^1 - 1) + k_i^{max})/t_i^1 \tag{30}$$

The total cost in [0, t_{i-1}^1] is $GAcost(t_{i-1}^1) = m_{i-1} * t_{i-1}^1$. The total cost in [0, $t_i^1 - 1$] is:

$$\begin{aligned} GAcost(t_i^1 - 1) &= t_i^1/t_{i-1}^1 * GAcost(t_{i-1}^1) - k_{i-1}^{max} \\ &= t_i^1/t_{i-1}^1 * m_{i-1} * t_{i-1}^1 - k_{i-1}^{max} \\ &= t_i^1 * m_{i-1} - k_{i-1}^{max} \end{aligned}$$

Replacing $GAcost(t_i^1-1)$ with $t_i^1*m_{i-1}-k_{i-1}^{max}$ in Equation (30),

$$m_i = (t_i^1*m_{i-1}+k_i^{max}-k_{i-1}^{max})/t_i^1 = m_{i-1}+(k_i^{max}-k_{i-1}^{max})/t_i^1$$

From Lemma 8.1, we have $t_i^1 \geq f_i/2$. Thus, we get

$$m_i \leq m_{i-1} + 2(k_i^{max} - k_{i-1}^{max})/f_i$$

□

Proof of Theorem 5.2

Proof: If we prove that $\forall t, 1 \leq t \leq f_i$,

$$m_i \leq 2*(DPcost(i, t-1) + k_i^{max})/t \tag{31}$$

then the theorem is proven. We use induction on i where $1 \leq i \leq g$. The basis for the induction, when $i = 1$, is to verify that inequality (31) holds. As discussed in Section 8, $m_1 = k_1^{max}/f_1$. We also know that $\forall t, 1 \leq t \leq f_1$, $DPcost(1, t-1) = 0$. Thus,

$$(DPcost(1, t-1) + k_1^{max})/t = k_1^{max}/t \geq k_1^{max}/f_1 = m_1$$

That is, $\forall t, 1 \leq t \leq f_1$, we know

$$m_1 \leq (DPcost(1,t-1)+k_1^{max})/t \leq 2*(DPcost(1,t-1)+k_1^{max})/t$$

For the induction step, suppose $\forall t, 1 \leq t \leq f_{i-1}$, we have $m_{i-1} \leq 2*(DPcost(i-1,t-1)+k_{i-1}^{max})/t$. That is, $\forall t, 1 \leq t \leq f_{i-1}$,

$$DPcost(i-1, t-1) + k_{i-1}^{max} \geq m_{i-1} * t/2 \tag{32}$$

As discussed earlier in Section 8, when $1 \leq t \leq f_i$, we have

$$DPcost(i, t-1) = DPcost(i-1, t-1) \tag{33}$$

If we want to prove $\forall t, 1 \leq t \leq f_i$,

$$m_i \leq 2*(DPcost(i, t-1) + k_i^{max})/t \tag{34}$$

we should prove

$$m_i \leq 2*(DPcost(i-1, t-1) + k_i^{max})/t \tag{35}$$

according to equation (33) and inequality (34). From the definition of $DPcost(i-1, t-1)$, we know

$$DPcost(i-1,t-1)+k_{i-1}^{max} = min(\sum DPcost(i-1,t_j-1)+k_{i-1}^{max}) \tag{36}$$

where $\forall t, 1 \leq t \leq f_i$, $\forall t_j, 1 \leq t_j \leq f_{i-1}$ and $\sum t_j = t$.

Replacing $DPcost(i-1, t-1)$ in inequality (35) with the formula in equation (36), our goal is to prove

$$m_i \leq 2*(min(\sum DPcost(i-1,t_j-1)+k_{i-1}^{max})-k_{i-1}^{max}+k_i^{max})/t. \tag{37}$$

According to inequality (32), we have

$$m_{i-1} * t_j/2 \leq DPcost(i-1, t_j - 1) + k_{i-1}^{max} \tag{38}$$

for each $1 \leq t_j \leq f_{i-1}$. Thus, if we want to prove inequality (37), we should prove

$$\begin{aligned} m_i &\leq 2*(min(\sum m_{i-1} * t_j/2) - k_{i-1}^{max} + k_i^{max})/t \\ &= 2*(m_{i-1} * t/2 - k_{i-1}^{max} + k_i^{max})/t \\ &= m_{i-1} + 2*(k_i^{max} - k_{i-1}^{max})/t \end{aligned} \tag{39}$$

As $1 \leq t \leq f_i$, we should prove

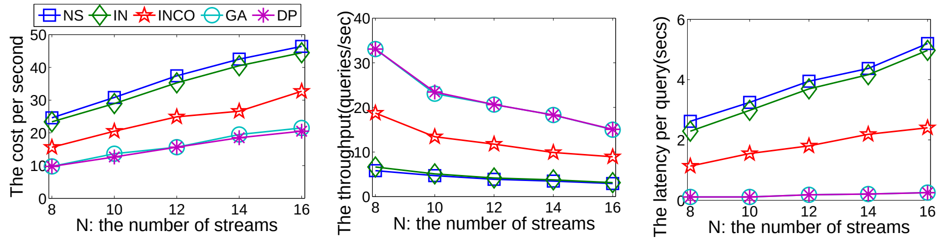
$$m_i \leq m_{i-1} + 2*(k_i^{max} - k_{i-1}^{max})/f_i \tag{40}$$

to achieve our original goal (34). From Theorem 8.1, Inequality (40) is proven. Then, our original goal (34) is achieved. That is, $\forall t, 1 \leq t \leq f_i$,

$$m_i \leq 2*(DPcost(i, t-1) + k_i^{max})/t$$

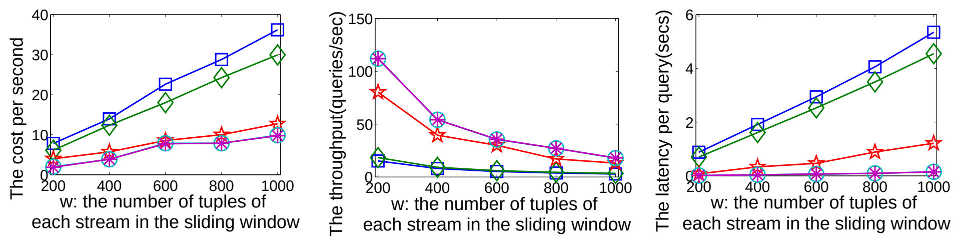
□

The results on real data sets for multiple streams are shown in Figure 19, Figure 20, Figure 21 and Figure 22. The trends are the same with those of synthetic data set for multiple streams.



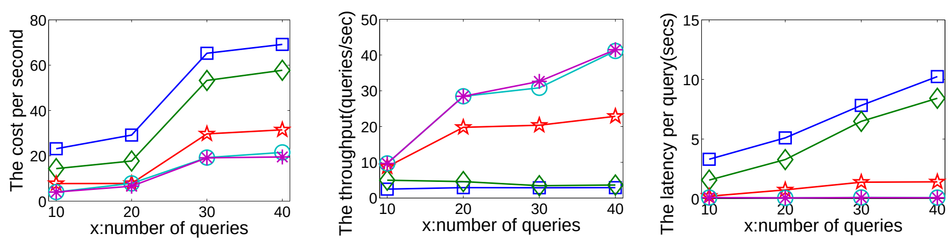
(a) The cost-per-second (b) The throughput (c) The latency

Fig. 19. The performance of DP, GA, INCO, IN and NS with different ranges of N



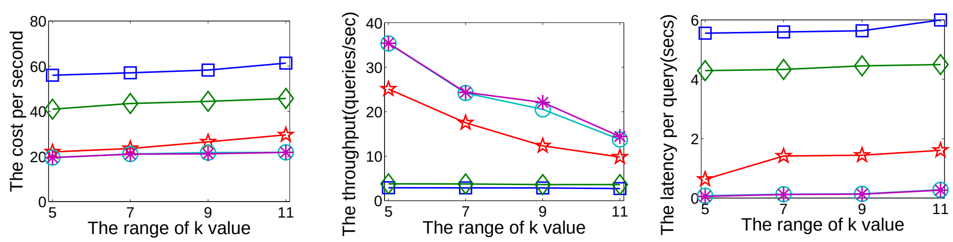
(a) The cost-per-second (b) The throughput (c) The latency

Fig. 20. The performance of DP, GA, INCO, IN and NS with different ranges of w



(a) The cost-per-second (b) The throughput (c) The latency

Fig. 21. The performance of DP, GA, INCO, IN and NS with different ranges of x



(a) The cost-per-second (b) The throughput (c) The latency

Fig. 22. The performance of DP, GA, INCO, IN and NS with different ranges of k