Approximate Range Thresholding

Anonymous Author(s)

ABSTRACT

In this paper, we study the (approximate) Range Thresholding (RT) problem over streams. Each element e in the stream is a point $v(e)$ in a d-dimensional space $\mathbb{U}^d$ and with a positive integer weight $w(e)$. An RT query q specifies an axis-parallel rectangular range $R(q)$ in $\mathbb{U}^d$ and a positive integer threshold τ. Once the query q is registered in the system, define $s(q)$ as the total weight of the elements that satisfy: (i) they arrive after q's registration, and (ii) they fall in the range $R(q)$. Given a real number $0 < \varepsilon < 1$, the task is to capture an arbitrary moment during the first moment when $s(q) \geq (1-\varepsilon)\tau$ and the first moment when $s(q) \geq \tau$. The challenge is to support a large number of RT queries simultaneously.

We propose a new algorithm called *FastRTS* , which can reduce the exponent in the poly-logarithmic factor of the state-of-the-art *QGT* algorithm from $d+1$ to d, yet slightly increasing the log term itself. A crucial technique to make this improvement happen is our *bucketing technique*, which eliminates the logarithmic factor caused by the use of heaps in *QGT* algorithm. Moreover, we propose two extremely effective optimization techniques which significantly improve the performance of *FastRTS* by orders of magnitude in terms of both running time and space consumption. We conduct extensive experiments on both synthetic and real datasets. Experimental results show that *FastRTS* outperforms the competitors by up to *three* orders of magnitude in both running time and peak memory usage.

ACM Reference Format:

Anonymous Author(s). 2021. Approximate Range Thresholding. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this paper, we study a type of queries, called *approximate range thresholding* queries (or *queries* for short when the context is clear). Specifically, we consider a system which receives *elements* from a *data stream* and supports this type of queries.

Stream Elements. In a data stream, each element e is a *point* $v(e)$ in a d-dimensional integer space $\mathbb{U}^d$, where $\mathbb{U} = \{1, 2, \ldots, N\}$ and d is a constant. Moreover, each element has an *integer weight* $w(e) > 0$.

Range Thresholding (RT) Queries. During the stream, an (approximate range thresholding) query q can be *registered* in the system. Each such query q specifies: (i) a *range* $R(q) = [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_d, b_d)$, which is a d-dimensional axis-parallel rectangular range in $\mathbb{U}^d$, where $a_i < b_i$ for all $i \in \{1, 2, \cdots, d\}$; and (ii) a positive integer *threshold* $\tau(q)$.

The Task of the System. For a query q, the task of the system is to "*track*" the total weight, denoted by $s(q)$, of the elements which: (i) arrive after the registration of the query q, and (ii) fall in the range $R(q)$. The system is required to *raise an alarm* (i.e., a notification) during the period between the *first moment* when $(1-\varepsilon)\cdot\tau(q) \leq s(q)$ and the *first moment* when $\tau(q) \leq s(q)$, for some specified parameter $0 < \varepsilon < 1$. Any alarm raised for q outside this period is deemed wrong. The goal of the system is to *correctly* raise an alarm for each of the queries that are registered in the system. We refer this problem as the *Range Thresholding* (RT) problem. A formal problem definition can be found in Section 2.

Applications. The range thresholding queries are especially useful in the scenarios where some time-critical actions must be taken as soon as certain events are detected. Here, the stock trading systems are good examples. In a stock trading system, each transaction of a stock (e.g., APPL:NSQ) can be considered as a stream element, where the selling (or buying) price is a 1-dimensional point and the number of shares traded in this transaction is the weight. For a query q, its range $R(q)$ can be a *sensitive* price range and the threshold $\tau(q)$ is a specified number of shares of the stock. As soon as a substantial total amount of shares (at least $(1-\varepsilon)\cdot\tau(q)$) are sold at sensitive prices in $R(q)$ since the registration of q, the system raises an alarm to notify the user. Such a notification is important because a substantial amount of shares sold at some sensitive prices is often an early signal of either a blow-up or a fall of the stock price. Thus, a timely notification is crucial for a user to take timely actions (e.g., either sell or buy the stock) to protect their interests.

Moreover, the above RT query in stock trading systems can be easily extended to a multi-dimensional one. For example, a 2-dimensional query may have a form like:

> *Notify me when in total $\tau(q)$ shares (from now) of Apple Inc. (APPL:NSQ) are sold at prices in* [\$140, \$150) *while the price of Alphabet Inc. (GOOG:NSQ) is in* [\$2800, \$2900).

Conventional Solutions. The RT problem is trivial if there is only one query. However, the challenge lies in supporting a large number of queries simultaneously. Let m be the number of queries and n be the number of elements arrived from the stream so far. For each element, this naive algorithm needs to check and increase counters (if necessary) for m queries. Hence, the time complexity becomes $O(n \cdot m)$, which is *quadratic* and thus, prohibitively expensive since both m and n can be very large.

More sophisticated solutions are to adopt certain data structures, such as the Interval Tree (for $d = 1$) [9, 13], the Segment Tree [13], and the R-tree [5, 17], to efficiently identify those queries whose ranges are "stabbed" by the element e, i.e., $v(e) \in R(q)$, and then increase their counters accordingly. However, since each query q can be stabbed as many as $(1-\varepsilon)\cdot\tau(q)$ times, the time complexity of these algorithms inevitably has the term $O((1-\varepsilon)\cdot\tau_{\max}\cdot m)$, which

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

is another form of *quadratic* term. Here, $\tau_{\max}$ is the largest threshold values among the queries. Unfortunately, none of the above conventional algorithms can overcome these quadratic bounds.

The State of the Art. The *QGT* algorithm [22] (named by the last names of its authors) is the first *sub-quadratic* solution for the RT problem. Its crucial idea is from an observation on the connection between the RT problem and a seemingly remote problem called the *Distributed Tracking* (DT) [10], where, interestingly, the latter problem is defined in a distributed environment. Based on the DT technique, the *QGT* algorithm achieves an overall time complexity of $O(m \cdot \log^{d+1} m \cdot \log \frac{1}{\varepsilon} + n \cdot \log^{d+1} m)$, and a space consumption bounded by $O(m_{\text{alive}} \cdot \log^d m_{\text{alive}})$ at all time, where m is the number of queries that have ever been registered in the system, n is the total number of elements from the stream so far, and m_{alive} is the number of queries that are still running in the system.

In theory, both the running time and the space consumption of the *QGT* algorithm are *near-linear* to both m and n, and hence, these bounds are considered as "efficient". However, in practice, just with $\log_2 m \approx 20$, the actual performance of the *QGT* algorithm are already significantly impacted by the *polynomial-logarithmic* factors with exponents $d+1$ (in the running time) and d (in the space) when $d \geq 3$. As we show later in the experiments (in Section 7), the performance of *QGT* deteriorates quickly and becomes worse than some of those aforementioned quadratic competitors for $d \geq 3$. Even worse, on 3-dimensional (rsp. 4-dimensional) datasets just with a moderate size of 2 million (rsp. 1 million) queries, the space consumption of *QGT* quickly exhausts all the available memory (100GB) of our machine! Hence, it fails to complete the corresponding experiment task. This space consumption issue has seriously limited the applicability of the *QGT* algorithm.

Our Contributions. We make the following contributions.

A New RT Algorithm. We propose a new algorithm called *FastRTS* for solving the RT problem. Specifically, our *FastRTS* achieves an expected overall running time complexity $O(m \cdot \log^d N \cdot \log \frac{1}{\varepsilon} + n \cdot \log^d N)$ and a space consumption bounded by $O(m_{\text{alive}} \cdot \log^d N)$ at all time, where N is the size of the universe $\mathbb{U}$ on each dimension. Comparing to the time complexity of the *QGT* algorithm, *FastRTS* reduces the exponential dependence on dimensionality d for the logarithmic term from $d+1$ to d, yet slightly increases this term from $\log m$ to $\log N$.

A Novel Bucketing Technique. We propose a new *Bucketing Technique* which is crucial for *FastRTS* to achieve the theoretical bounds. As we will see in Section 5.2, our bucketing technique allows *FastRTS* to eliminate the $O(\log m)$-factor caused by the use of min-heaps for organizing the DT instances. We note that this bucketing technique is of *independent interests*; it may be able to remove the similar logarithmic factor overhead for all those algorithms which need to organize DT instances with heaps, e.g., the *QGT* algorithm and a very recent work [23].

A New DT Algorithm. In order to facilitate our bucketing technique, we propose a new DT algorithm called *Power-of-Two-Slack DT* (P2S-DT) in Section 5.1. We perform theoretical analysis to show both the correctness and the communication cost of the algorithm. A nice property of our P2S-DT is that all the slack values in this algorithm are power-of-two integers. Thus, our P2S-DT may be of interests in certain scenarios.

Two Effective Optimizations. It should be noted that the space consumption bound of our *FastRTS* is slightly worse than that of the *QGT* algorithm. To remedy this, we propose two *extremely effective* optimization techniques: (i) the *Range Shrinking* technique (Section 6.1), and (ii) the *Range Counting* technique (Section 6.2). We show that these two techniques not only theoretically sound, but also dramatically improve the actual performance of *FastRTS* . According to the experimental results in the ablation study, these two techniques improve both the running time and space consumption of our *FastRTS* by up to *three* orders of magnitude.

Extensive Experiments. We conduct extensive experiments on both synthetic datasets and two real stock trading datasets. Experimental results show that our *FastRTS* outperforms all the state-of-the-art competitors by orders of magnitude in both overall running time and space consumption.

2 PRELIMINARIES

In this section, we first formally define the *Range Thresholding* (RT) problem and then introduce the *Distributed Tracking* technique and the state-of-the-art *QGT* algorithm for solving the RT problem. These background knowledge will ease the understanding on our proposed techniques.

2.1 The RT Problem Formulation

Let $\mathbb{U} = \{1, 2, \ldots, N\}$ be a finite consecutive integer domain, whose size is $|\mathbb{U}| = N$, and where each integer in $\mathbb{U}$ can be encoded with $O(1)$ *words*[1]. All the integers considered in this paper are upper bounded by a polynomial of N and hence, they are all $O(1)$-word representable.

Consider a system which receives data *elements* from a *stream* S and supports the *range thresholding queries*.

The Stream and Elements. The stream S is a sequence of elements $e_1, e_2, \cdots$; the i-th element e_i arrives at *time stamp* i, where $i = 1, 2, \cdots$. In particular, at time stamp 0, the stream S is empty: no element has arrived yet. Each element e_i consists of two fields:

- a *value*, denoted by $v(e_i)$, which is a d-dimensional point in $\mathbb{U}^d$;
- a *weight*, denoted by $w(e_i)$, which is a *positive integer*.

Range Thresholding Queries. Each (range thresholding) query q can be registered in the system at any time stamp *immediately after* the arrival of the corresponding element. Specifically, a query is composed of two fields:

- a *range*, denoted by $R(q) = [a_1, b_1) \times [a_2, b_2) \times \cdots [a_d, b_d)$, which is a d-dimensional axis-parallel rectangular range in $\mathbb{U}^d$, where a_i and b_i are integers and $a_i < b_i$ for all $i = 1, 2, \cdots, d$;
- a *threshold*, denoted by $\tau(q)$, which is a *positive integer*.

ε-Maturity. Consider a query q *registered* in the system at time stamp t; for any time stamp $t' > t$, define $S(q, t, t') \subseteq S$ as the set of all the elements which (i) arrive between time stamp $t+1$ and t', and (ii) have values falling within the range $R(q)$. That is,

$$S(q, t, t') = \{e_i \in S \mid v(e_i) \in R(q) \wedge t < i \leq t'\}.$$

[1]Throughout this paper, space consumptions are measured by number of words.

Given a query q and a real number $0 < \varepsilon < 1$, the task of the system is to *raise an alarm* during the period of time between the *first moment* t' when $(1-\varepsilon)\cdot\tau(q) \le \sum_{e\in S(q,t,t')} w(e)$, and the *first moment* t^* when $\tau(q) \le \sum_{e\in S(q,t,t^*)} w(e)$. Such a time period of $[t', t^*]$ is called *ε-maturity period* of q. Any time stamp $\ell \in [t', t^*]$ is defined as the *ε-maturity* of q.

Definition 2.1 (The Range Thresholding (RT) Problem). Given a stream S of elements and a real number $0 < \varepsilon < 1$, the *Range Thresholding problem* is to design an algorithm for the system such that: (i) it supports dynamic query registrations and terminations, and (ii) it correctly captures an *arbitrary* one of the ε-maturity moments for every alive query.

About ε. Although in Definition 2.1, the parameter ε is specified to be the same for all queries, each query can actually have a dedicated constant ε value. In particular, by setting $\varepsilon < 1/\tau(q)$, the system can *exactly* capture the first moment t^* for the query q.

Loosely Related Work. We hereby describe a range of (sub-)fields that are loosely related to our problem.

The first one is *triggers* in DBMS, which can be viewed as a form of RT in concept. Triggers were primarily designed to ensure certain integrity constraints on underlying relations. Subsequently, more complicated forms of triggers were introduced to enable a DBMS to activate itself in response to a greater variety of events, and many efforts have been made to implement such triggers in an efficient manner, falling in the topic of active databases [21].

The second field within which the RT problem might fall is *continuous query processing* over data streams [2, 3, 8, 19]. Essentially, a continuous query is a query that is issued once over a database, and then logically runs continuously over the data in database until it is terminated. When mapping it to the RT problem, the user issuing an RT query should get an alert at the query's maturity time.

The third sub-field is a variant of triggers on streaming data, called *publish/subscribe system*. In this context, users can specify their particular interest via a subscription query (e.g. keywords, tags), such that whenever new data elements flow through the system (e.g. tweets, products, promotions, news articles, etc.), only those elements that are "relevant" to a user's subscription will be "pushed" to the user instantly. In this way, users can alleviate themselves from being overwhelmed in the era of information explosion. Interested readers can refer to [15, 16, 18, 20, 25] for some representative work. To this end, we find RT can be viewed as a form of subscription but even so, it is a new type of subscription query with its unique computational challenge.

The last sub-field is essentially a rejuvenation of the aforementioned active databases on data streams, which yields another line of research under the umbrella of complex event processing [24]. We refer interested readers to some earlier work [12, 14] for more background knowledge. Our work actually complements it in the sense that RT can be treated as another atomic event type that needs to be supported in an efficient and scalable manner.

2.2 Distributed Tracking

The DT problem [11] is defined in a distributed environment, where there are h *participants* u_i (for $i = 1, 2, \cdots, h$) and one *coordinator* q. Each participant can only communicate with the coordinator q by sending and receiving *messages*, each with $O(1)$ *words*. For each participant u_i, there is a *counter* c_i which is 0 initially. The coordinator q has a positive integer *threshold* τ. At each time stamp, *at most* one (that means, it could be none but no more than one) participant has its counter increased by an arbitrary positive integer value. Given a real number $0 < \varepsilon < 1$, the task of the coordinator is to capture an *arbitrary moment* during the *ε-maturity period* which is the period between the *first moment* when $\sum_{i=1}^{h} c_i \ge (1-\varepsilon)\cdot\tau$ and the *first moment* when $\sum_{i=1}^{h} c_i \ge \tau$. Any such moment during the maturity period is called a *ε-maturity moment*. The goal is to minimize the communication cost: the total number of messages sent and received by the coordinator.

A straightforward solution is to instruct each participant to send a message to the coordinator q when its counter is increased. In this way, q can keep track of the counter sum precisely and report ε-maturity as soon as this sum $\ge (1-\varepsilon)\cdot\tau$. Clearly, the communication cost is $O((1-\varepsilon)\cdot\tau)$ as each counter increment could be 1. When the threshold τ is large, the $O((1-\varepsilon)\cdot\tau)$ communication cost is expensive. A state-of-the-art DT algorithm [10] can achieve a $O(h\cdot\log\frac{1}{\varepsilon})$ communication cost bound. Let $\bar{\tau}$ be the current threshold and initially, $\bar{\tau} \leftarrow \tau$. The DT algorithm works in *rounds*; in each round, it works as follows:

- If $\bar{\tau} \le 2h$, run the naive algorithm with $O(\bar{\tau}) = O(h)$ communication cost and done;
- Otherwise,
  - q sends a *slack* $\lambda = \lfloor \frac{\bar{\tau}}{2h} \rfloor$ to each participant;
  - for every λ increments on the counter c_i, participant u_i sends a message to q;
  - when q receives the h^{th} message in the current round, q collects the precise counters of all the participants and computes $\tau' = \bar{\tau} - \sum_{i=1}^{h} c_i$; if $\tau' \le \varepsilon\cdot\tau$, q reports ε-maturity; otherwise, q starts a new round by running a new DT instance with threshold $\bar{\tau} \leftarrow \tau'$ from scratch (with all the counters c_i reset to 0).

Analysis. It can be verified that at the end of each round, τ' is at most a constant fraction of the threshold $\bar{\tau}$, and no more rounds will be performed when $\bar{\tau} \le 2h$. As a result, there are $O(\log\frac{\tau}{\varepsilon\cdot\tau}) = O(\log\frac{1}{\varepsilon})$ rounds. In each round, the coordinator q receives and sends at most $O(h)$ messages. The total communication cost is thus bounded by $O(h\cdot\log\frac{1}{\varepsilon})$. We summarize this result in the following fact:

Fact 1 ([11]). *There exists an algorithm which solves the DT problem with $O(h\cdot\log\frac{1}{\varepsilon})$ communication cost.*

2.3 The State-of-the-Art RT Algorithm

Next, we introduce the *QGT* algorithm [22], a state of the art for solving the RT problem.

The One-Time Registration Case. For the ease of discussion, we start with a *restricted case* that all the m queries are registered before the first element in the stream S arrives, i.e., at time stamp 0, and since then, no query registrations are allowed. We refer this case as the *one-time registration case*.

The d-Dimensional Endpoint Tree. The basic idea of the *QGT* algorithm is to construct a d-dimensional Segment Tree [13] on all the endpoints of the m queries. Such a Segment Tree is called the

Endpoint Tree and denoted by $\mathcal{T}$. It is known that there are d *layers* of base trees in $\mathcal{T}$, where the i^{th}-layer tree is constructed on the i^{th} dimension. The tree nodes in the d^{th}-layer trees are called *last-layer nodes*. Each last-layer node u is associated with a *range* $R(u) \subseteq \mathbb{U}^d$ and a *counter* c_u which records the sum of the weights of the elements in the stream S that fall in $R(u)$. Moreover, each such last-layer node u also contains a list of queries, denoted by $L(u)$, which have u in their *canonical node sets*. The latter notion is defined as follows.

Canonical Node Set. With the Endpoint Tree $\mathcal{T}$, the range $R(q)$ of each query q can be decomposed into a set of *canonical sub-ranges*, denoted by $C(q)$, such that: (i) each sub-range is the associated range of some last-layer node u in $\mathcal{T}$, (ii) the sub-ranges in $C(q)$ are all *disjoint* and their union is *precisely* equal to $R(q)$, and (iii) $C(q)$ is the smallest (in terms of size) among all possible sub-range sets satisfying (i) and (ii). The set of all the last-layer nodes u whose ranges are in $C(q)$ is defined as the *canonical node set* of q, denoted by $\mathcal{N}(q)$. Therefore, the total weight of those elements that fall in $R(q)$ is equal to the counter sum of all the nodes in $\mathcal{N}(q)$, namely,

$$\sum_{u \in \mathcal{N}(q)} c_u = \sum_{u \in \mathcal{N}(q)} \left(\sum_{e \in S \wedge v(e) \in R(u)} w(e) \right) = \sum_{e \in S \wedge v(e) \in R(q)} w(e) .$$

Creating DT Instances. The query q and its canonical node set $\mathcal{N}(q)$, in fact, constitute a DT instance, denoted by $DT(q)$. Specifically, the query q is the coordinator with $\tau(q)$ as the threshold and each node $u \in \mathcal{N}(q)$ is a participant with c_u as the counter. Therefore, an ε-maturity moment of q can be captured by $DT(q)$.

Organizing DT Participants with Heaps. Observe that if $L(u)$ is non-empty, each last-layer node u works as a participant in the DT instance $DT(q)$ for each $q \in L(u)$. While the counter c_u is *shared* by all these DT instances, the participants that u is playing in different instances may need to "communicate" with their corresponding coordinators (i.e., queries) at different time. For every increment of c_u, it is costly to check all these participants of u if they need to notify the coordinator. Worse still, it ends up results in an overall quadratic cost $O((1-\varepsilon) \cdot \tau_{\max} \cdot m)$.

To address this issue, the participants of u in the $DT(q)$ for all $q \in L(u)$ are maintained by a *min-heap*, denoted by $\mathcal{H}(u)$, with a *priority* calculated as $\lambda_q + \bar{c}_q$, where λ_q is the current slack value in $DT(q)$, and $\bar{c}_q$ is the value of c_u when this participant *last* sent a message to the coordinator q (initially, $\bar{c}_q = 0$). Specifically, this priority is essentially the *check-point* counter value indicating that when c_u reaches to this value, the corresponding participant needs to send a message to the coordinator. Therefore, when c_u is increased, it suffices to check whether $c_u \geq \mathcal{H}(u).top$, where $\mathcal{H}(u).top$ is the *smallest* priority value in the heap, i.e., the smallest check-point counter value. If so, the participant of u (say, in the DT instance $DT(q)$) at the top in $\mathcal{H}(u)$ is extracted. Following the steps in the DT algorithm, this participant sends a message to its coordinator and updates its priority accordingly (because $\bar{c}_q$ becomes the current counter value of u and λ_q may also change due to the start of a new round). The heap top is repeatedly extracted until no more participants need to notify their coordinators (i.e., $c_u < \mathcal{H}(u).top$). Moreover, when $DT(q)$ ends its current round, it suffices to update the priorities of all the participants in $DT(q)$ accordingly in the heap $\mathcal{H}(u)$ for each $u \in \mathcal{N}(q)$.

Running Time Analysis. To analyse the overall running time, we consider two types of costs: (i) the cost for element processing, and (ii) the cost for the DT instances.

First, for each element e, the *QGT* algorithm needs to increase the counter c_u by $w(e)$ for all the last-layer nodes u whose ranges are "*stabbed*" by e, namely, $v(e) \in R(u)$. This can be achieved by a standard *stabbing* query on the Endpoint Tree $\mathcal{T}$. It is known that each such query takes $O(\log^d m)$ time. Thus, the total element processing cost is bounded by $O(n \cdot \log^d m)$.

Second, for the DT part, by implementing the min-heap of a node u with the standard binary heap [9], each heap operation can be performed in $O(\log |L(u)|) = O(\log m)$ time. Therefore, the cost of each $DT(q)$ takes $O(|\mathcal{N}(q)| \cdot \log \frac{1}{\varepsilon} \cdot \log m)$ time. Furthermore, by the standard result of the Segment Tree [13], the cardinality of the canonical node set, $|\mathcal{N}(q)|$, is bounded by $O(\log^d m)$. Therefore, the cost for each $DT(q)$ is bounded by $O(\log^{d+1} m \cdot \log \frac{1}{\varepsilon})$, and thus, the overall cost is bounded by $O(m \cdot \log^{d+1} m \cdot \log \frac{1}{\varepsilon})$ for m queries.

Putting the above two costs together gives the overall running time cost for the one-time registration case, which is bounded by $O(m \cdot \log^{d+1} m \cdot \log \frac{1}{\varepsilon} + n \cdot \log^d m)$.

Space Analysis. The space consumption is dominated by the size of the Endpoint Tree $\mathcal{T}$ along with all the DT instances. Thus, this size is bounded by $O(m \cdot \log^d m)$. Moreover, by a careful tree rebuilding strategy (i.e., rebuilding everything if $m_{\text{alive}} < m/2$), the *QGT* algorithm warrants a space consumption bounded by $O(m_{\text{alive}} \cdot \log^d m_{\text{alive}})$ *at all time*, where m_{alive} is the number of queries that are alive in the system.

Supporting Query Dynamics. In order to support query registrations and terminations (referred as *query dynamics*), the *QGT* algorithm adopts the standard *logarithmic method* [4, 6], which is a general technique for making a static data structure dynamic. In the case of *QGT* algorithm, the logarithmic method introduces an extra logarithmic factor, i.e., $O(\log m)$, to the element processing cost, and admits an amortized cost of $O(\log^{d+1} m)$ for each query registration and termination, while keeping the space consumption bound unchanged. The final result is summarized below:

Fact 2. *For m queries (along with their registrations and terminations) and n elements, the* QGT *algorithm solves the ε-approximate RT problem with:*

- $O(m \cdot \log^{d+1} m \cdot \log \frac{1}{\varepsilon} + n \cdot \log^{d+1} m)$ *overall running time, and*
- $O(m_{alive} \cdot \log^d m_{alive})$ *space consumption at all time.*

3 AN OVERVIEW OF OUR ALGORITHM

In this section, we give an overview of our proposed algorithm, called *Fast Range Thresholding over Streams* (*FastRTS*), for solving the RT problem.

The Algorithmic Framework. *FastRTS* adopts the same algorithmic framework of *QGT* . Specifically, that is to:

- maintain a certain tree-based data structure (details are in Section 4), which decomposes each query range $R(q)$ into a set of

canonical sub-ranges $C(q)$, and the set of the corresponding tree nodes of the ranges in $C(q)$ is the canonical node set $\mathcal{N}(q)$;
- create a DT instance $DT(q)$ with the nodes in $\mathcal{N}(q)$ as participants for each query q, to capture its ε-maturity;
- for each last-layer tree node u, organize the DT instances where u serves as a participant in a certain way (details are in Section 5) such that these DT instances can be run simultaneously in an efficient manner.

The *FastRTS* Algorithm. There are two main modules in *FastRTS*: (i) the Query and Element Processing (QEP) Module, and (ii) the DT Manager (DTMgr) Module.

Query and Element Processing (QEP) Module. This module maintains an incremental Segment Tree (*IncSegTree*, in Section 4) such that:

- for the registration of each query q, decompose the range $R(q)$ into $C(q)$ and associate q to the nodes in $\mathcal{N}(q)$, and then invoke DTMgr module's procedure for query registrations;
- for the termination of each query q, invoke DTMgr module's procedure for query terminations, and then remove q from the *IncSegTree*;
- for each element e, increase the counters of those last-layer nodes whose ranges contain the point $v(e)$ by weight $w(e)$, and then invoke DTMgr module's procedure for counter increments for each of these nodes.

DT Manager (DTMgr) Module. The DTMgr module, for each last-layer node u, organizes the participants of u (sometimes, we just refer these participants to the corresponding queries) in all the DT instances $DT(q)$ for $q \in L(u)$, where $L(u)$ is the set of all the queries which have u in $\mathcal{N}(q)$, with our proposed *Bucketing Technique* (in Section 5.2). To facilitate this technique, all the DT instances are run with our new variant of the DT algorithm, called *Power-of-Two-Slack DT* (P2S-DT) algorithm (in Section 5.1). There are three main interface procedures in this DTMgr module.

- Procedure for Query Registrations. It creates a DT instance for a query q by inserting q to the *bucketing structure* in each node $u \in \mathcal{N}(q)$.
- Procedure for Query Terminations and Maturity. It removes the query q from the bucketing structure in each node $u \in \mathcal{N}(q)$.
- Procedure for Counter Increments. This procedure manages the DT instances following the P2S-DT algorithm for each node u whose counter is increased.

In the next two sections, we show the details of the above two modules and prove the following theorem:

Theorem 3.1. *For m queries (along with their registrations and terminations) and n elements, the* FastRTS *algorithm solves the ε-approximate RT problem with:*

- $O(m \cdot \log^d N \cdot \log \frac{1}{\varepsilon} + n \cdot \log^d N)$ *overall running time in expectation,*
- $O(m_{alive} \cdot \log^d N)$ *space consumption at all time,*

where N is the size of the universe $\mathbb{U}$ on each dimension.

4 THE QEP MODULE

In this section, we reveal the details of the Query and Element Processing module in *FastRTS*. Unlike the *QGT* algorithm, which maintains a set of *Endpoint Trees* with the logarithmic method, our QEP module adopts an *Incremental Segment Tree* (*IncSegTree*) on the universe $\mathbb{U}^d$ to handle query registrations and terminations.

A Full Segment Tree on $\mathbb{U}^d$. For the ease of presentation, we first start with a simple idea without worrying about the space consumption. This idea is to construct a *full* Segment Tree on the universe $\mathbb{U}^d$, where the base tree in each layer is on $\mathbb{U}$ of the corresponding dimension. Specifically, we build a Segment Tree on $\mathbb{U}$ for the first dimension, and then for each node, we recursively build a Segment Tree on $\mathbb{U}$ for the next dimension.

Example 1. *Figure 1(a) shows a full Segment Tree $\mathcal{T}$ (including all the grey nodes) on $\mathbb{U} = \{1, \ldots, 16\}$. Furthermore, for the 2-dimensional case, each node of $\mathcal{T}$ is associated a full Segment Tree on the next dimension as shown in Figure 1(b).*

As the full Segment Tree captures all the possible endpoints on each dimension, the structure of the tree is *static* (i.e., fixed) under query insertions and deletions, namely, there is no need to add or remove endpoints from the tree. Therefore, a query insertion (deletion) is just as simple as to associate (remove) the query to (from) the corresponding canonical nodes in the tree, which, by the standard results of Segment Tree [13, 22], can be performed in $O(\log^d N)$ time. Furthermore, for each element e, the counters of all those last-layer nodes, whose associated ranges contain $v(e)$, can be increased in $O(\log^d N)$ time.

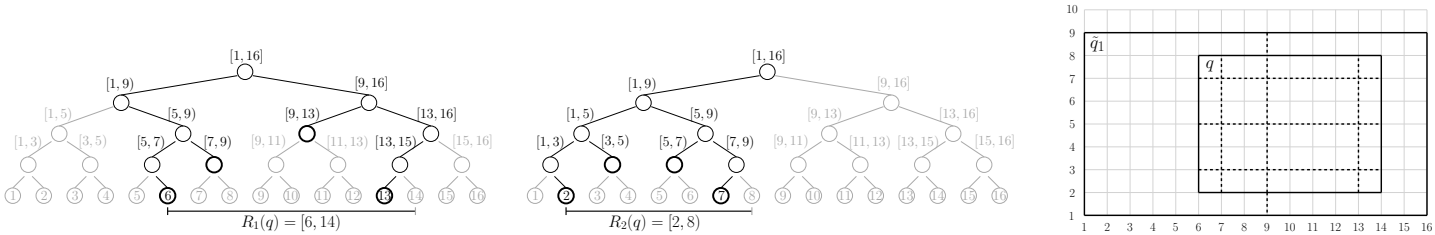
Example 2. *Continuing the previous example, as shown in Figure 1, a query q with $R(q) = [6, 14) \times [2, 8)$ is decomposed by the Segment Tree on the first dimension into four nodes with ranges $R_1(q) = [6, 7) \cup [7, 9) \cup [9, 13) \cup [13, 14)$, and by the Segment Tree on the second dimension into four nodes such that $R_2(q) = [2, 3) \cup [3, 5) \cup [5, 7) \cup [7, 8)$ shown in Figure 1(b). Thus, $R(q)$ is decomposed into $4 \times 4 = 16$ sub-ranges as shown in Figure 1(c).*

While the full Segment Tree is efficient for query dynamics, its space consumption is an issue. To see this, there are $O(N)$ nodes in the base tree at the first layer of a full Segment Tree on $\mathbb{U}^d$; each of these nodes has a full Segment Tree on $\mathbb{U}^{d-1}$. As a result, solving the recursion gives the overall space of the d-dimensional base tree bounded by $O(N^d)$. This space consumption is prohibitive.

An Incremental Segment Tree on $\mathbb{U}^d$. We address this issue by constructing a full Segment Tree on $\mathbb{U}^d$ *incrementally*. This variant is called an *Incremental Segment Tree* (*IncSegTree*). More specifically, the basic idea of the *IncSegTree* is to perform query insertions and deletions as if it is a full Segment Tree, yet a node in *IncSegTree* is materialized only when it is "touched" by some query. For example, as shown in Figure 1, the nodes in black are materialized while those in grey are not. From the standard result [13, 22], a query insertion and deletion can only touch at most $O(\log^d N)$ nodes in a full Segment Tree. Thus, only the same number of nodes in *IncSegTree* will be materialized. Thus, for m query insertions, the space consumption of the *IncSegTree* is bounded by $O(m \cdot \log^d N)$. For query deletions, a node in *IncSegTree* is deleted only when there is no query associated to any node in its sub-tree. We summarize the key results in the following fact.

Fact 3. *For a d-dimensional Incremental Segment Tree on $\mathbb{U}^d$,*

- *each query insertion and deletion takes in $O(\log^d N)$ time;*



(a) *IncSegTree* materialized by q on the first dimension (b) *IncSegTree* materialized by q on the second dimension (c) Canonical Sub-ranges of $R(q)$ and of $R(\tilde{q}_1)$

Figure 1: An *IncSegTree* on $\mathbb{U}^2 = \{1, \ldots, 16\}^2$ for a query q with $R(q) = [6, 14) \times [2, 8)$

- *for each query q, $|\mathcal{N}(q)| = O(\log^d N)$;*
- *each element from the stream can be processed in $O(\log^d N)$ time;*
- *the space consumption is bounded by $O(m_{alive} \cdot \log^d N)$ at all time.*

5 THE DT MANAGER MODULE

In this section, we first introduce a new variant of the DT algorithm, namely *Power-of-Two-Slack DT* (P2S-DT). A nice property of the P2S-DT is that the slack values are all power-of-two integers. With the P2S-DT algorithm, we introduce our *bucketing technique* to organize the DT instances. For each last-layer node u, instead of using a min-heap, the bucketing technique adopts a linked list of *non-empty* buckets, where each bucket stores all the queries in $L(u)$ having the same power-of-two slacks. As we will show in Section 5.2, our bucketing technique can eliminate the $O(\log m)$ factor caused by the min-heaps from the overall DT cost.

5.1 A New DT Algorithm

To facilitate our bucketing technique, we first develop a new DT algorithm, called *Power-of-Two-Slack DT* (P2S-DT). The P2S-DT algorithm works in rounds. Let $\bar{\tau}$ be the threshold in the current round; initially, $\bar{\tau} \leftarrow \tau$. In each round it works as follows:

- If $\bar{\tau} \leq 2h$, run the straightforward algorithm with communication cost $O(\bar{\tau}) = O(h)$ and done.
- Otherwise,
  - record $c'_i = c_i$, the *initial counter value* of participant u_i at the start of the current round, for $i = 1, 2, \cdots, h$;
  - let $\bar{c}_i$ be the counter value when participant u_i communicated with q last time; initially, $\bar{c}_i = c'_i$;
  - the coordinator q sends a slack $\lambda = \lfloor \frac{\bar{\tau}}{2h} \rfloor_2$ to each of the participants, where the operation $\lfloor x \rfloor_2$ returns the *largest power-of-two integer* which is *no larger* than x, that is, $\lfloor x \rfloor_2 = 2^{\lfloor \log_2 x \rfloor}$;
  - define *λ-multiple successor* of an integer x as the *smallest multiple of λ* which is *no less* than x, i.e., $\lambda \cdot \lceil \frac{x}{\lambda} \rceil$;
  - when the counter c_i of a participant u_i *reaches or exceeds* the λ-multiple successor of $\bar{c}_i$, i.e., $c_i \geq \lambda \cdot \lceil \frac{\bar{c}_i}{\lambda} \rceil$, u_i sends to q the counter increment since last report, i.e., $c_i - \bar{c}_i$, and then updates $\bar{c}_i$ to c_i, i.e., $\bar{c}_i \leftarrow c_i$;
  - when the total counter increment that q received since the current round is $\geq \frac{\bar{\tau}}{2}$, q collects the precise counters from all the participants and calculates $\tau' = \bar{\tau} - (\sum_{i=1}^h c_i - \sum_{i=1}^h c'_i)$; if $\tau' \leq \varepsilon \cdot \tau$, q reports the maturity; otherwise, start a new round with threshold $\bar{\tau} \leftarrow \tau'$ and with all the current counter values *unchanged* (i.e., the counter values are not reset).

Correctness and the Communication Cost. Next, we prove the correctness of the above P2S-DT algorithm and then analyse its communication cost.

Lemma 5.1. *During a round before q receives $\bar{\tau}/2$ counter increment, the total counter increment must satisfy: $\sum_{i=1}^h c_i - \sum_{i=1}^h c'_i < \bar{\tau}$, and hence, q will not miss the ε-maturity period.*

Proof. According to the P2S-DT algorithm, at any time during a round before q receives $\bar{\tau}/2$ counter increment, the total *actual* counter increment is upper bounded by the sum of the counter increment received by q and the maximum possible total counter increment that q is not yet aware of, i.e., $h \cdot (\lambda - 1)$. Thus,

$$\sum_{i=1}^h c_i - \sum_{i=1}^h c'_i < \frac{\bar{\tau}}{2} + h \cdot (\lambda - 1) < \frac{\bar{\tau}}{2} + h \cdot \frac{\bar{\tau}}{2h} = \bar{\tau} .$$

□

Lemma 5.2. *In each round, the coordinator receives at most $3h$ messages from the participants.*

Proof. It suffices to show that after receiving $3h$ messages from the participants, q must be aware of a total counter increment $\geq \bar{\tau}/2$ in the current round.

Recall that in P2S-DT algorithm, a participant u_i sends a counter increment to q only when c_i reaches or exceeds the λ-multiple successor of $\bar{c}_i$. As a result, in the worst case, the counter increment $c_i - \bar{c}_i$ can be as small as just 1. However, this worst case can only happen for the *first* report of each participant. It can be verified that if a participant u_i has reported $k + 1$ times (for $k \geq 0$), the total counter increment in u_i that q is aware of is at least $k \cdot \lambda + 1$. Therefore, for $3h$ reports from the participants, in the worst case, the total counter increment which q is aware of is at least:

$$h \cdot 1 + 2 \cdot h \cdot \lambda = h + 2 \cdot h \cdot \lfloor \frac{\bar{\tau}}{2h} \rfloor_2 \geq h + 2 \cdot h \cdot \frac{\bar{\tau}}{4h} \geq \frac{\bar{\tau}}{2} .$$

□

By Lemma 5.2, we know that there are at most $O(\log \frac{1}{\varepsilon})$ rounds, each of which takes $O(h)$ communication cost. Combining Lemma 5.1, we have the following theorem:

Theorem 5.3. *The P2S-DT correctly solves the ε-approximate DT problem with $O(h \cdot \log \frac{1}{\varepsilon})$ communications. When $\varepsilon < \frac{1}{\tau}$, it captures the exact maturity with $O(h \cdot \log \frac{\tau}{h})$ communications.*

5.2 The Bucketing Technique

Let us go back to the context of the RT problem. According to the algorithmic framework of both our *FastRTS* and the *QGT* algorithms, for a last-layer node u in the d-dimensional (incremental) Segment Tree, u serves as a participant in those DT instances $DT(q)$'s, for all $q \in L(u)$, where $L(u)$ is the list of all the queries q having u in the canonical node set $\mathcal{N}(q)$. When u's counter is incremented, u needs to identify all those DT instances, in which the participant of u should notify the corresponding coordinators according to the DT algorithm. In the following, the DT algorithm we consider is the P2S-DT algorithm, where all the slack values are power-of-two integers.

A Linked List of Buckets. As mentioned earlier, for a last-layer node u, we organize $L(u)$ with a *linked list* of buckets. Specifically, for each $i \in \{0, 1, \ldots, \lfloor \log_2 \tau_{\max} \rfloor\}$, we *conceptually* maintain a dedicated bucket, denoted by $\mathcal{B}_i$, to store all the queries $q \in L(u)$ where the participants of u have a slack value of 2^i in the corresponding DT instances. However, physically maintaining all these buckets for each last-layer node u may cause an extra $O(\log N)$ overhead in the overall space consumption. Instead, for each last-layer node u, we only maintain a linked list, denoted by $\mathcal{A}(u)$, of all those *non-empty* buckets sorted in ascending order by their corresponding slack values. Furthermore, for each bucket $\mathcal{B}$, we maintain a counter, denoted by $\bar{c}_u(\mathcal{B})$, to record the value of c_u when the queries in $\mathcal{B}$ last notified their coordinators; initially, $\bar{c}_u(\mathcal{B})$ is set to the value of c_u when $\mathcal{B}$ is created. Similarly, we also maintain a counter, denoted by $\bar{c}_u(q)$, for each query q to record the value of c_u when the participant of $DT(q)$ last communicated with the coordinator; initially, $\bar{c}_u(q)$ is set to the value of c_u when q is registered.

The Detailed Procedures of the Bucketing Technique.

Procedure for Query Registrations. When a query q is registered, for each node $u \in \mathcal{N}(q)$, perform the following steps:

- find the bucket $\mathcal{B}$ corresponding to the slack of $DT(q)$ in $\mathcal{A}(u)$;
- if $\mathcal{B}$ does not exist in $\mathcal{A}(u)$,
  - create the bucket $\mathcal{B}$ and insert $\mathcal{B}$ at a *proper position* in $\mathcal{A}(u)$ such that all the buckets are still sorted;
  - set $\bar{c}_u(\mathcal{B}) \leftarrow c_u$;
- set $\bar{c}_u(q) \leftarrow c_u$, and store q in $\mathcal{B}$;

Procedure for Query Terminations and Maturity. When a query q is terminated or matured, for each node $u \in \mathcal{N}(q)$,

- locate q in the $\mathcal{B}$ in $\mathcal{A}(u)$;
- remove q from $\mathcal{B}$;
- if $\mathcal{B}$ becomes empty, remove $\mathcal{B}$ from $\mathcal{A}(u)$.

Procedure for New Round Starting. When a DT instance $DT(q)$ starts a new round with a *new* slack $\lambda = 2^i$, for each node $u \in \mathcal{N}(q)$, perform the following steps:

- locate q in the bucket $\mathcal{B} \in \mathcal{A}(u)$;
- traverse the linked list $\mathcal{A}(u)$ from $\mathcal{B}$ towards the *head* of $\mathcal{A}(u)$ until one of the following three cases holds:
  - Case 1. The bucket $\mathcal{B}_i$ corresponding to λ is found;
  - Case 2. A bucket $\mathcal{B}'$ of slack $< \lambda$ is met: create the bucket $\mathcal{B}_i$; insert $\mathcal{B}_i$ after $\mathcal{B}'$ in $\mathcal{A}(u)$; set $\bar{c}_u(\mathcal{B}_i) \leftarrow c_u$;
  - Case 3. The head $\mathcal{B}'$ of $\mathcal{A}(u)$ is of slack $> \lambda$: create the bucket $\mathcal{B}_i$; insert $\mathcal{B}_i$ before $\mathcal{B}'$ in $\mathcal{A}(u)$; set $\bar{c}_u(\mathcal{B}_i) \leftarrow c_u$;
- remove q from $\mathcal{B}$, add q to $\mathcal{B}_i$, and set $\bar{c}_u(q) \leftarrow c_u$;
- if $\mathcal{B}$ becomes empty, remove $\mathcal{B}$ from $\mathcal{A}(u)$;

Procedure for Counter Increments. When c_u is increased, perform the following steps:

- let $\mathcal{B}$ be the current bucket in $\mathcal{A}(u)$ of a slack value λ; initially, $\mathcal{B}$ is set as the first bucket (i.e., the head) of $\mathcal{A}(u)$;
- if $c_u \geq \lambda \cdot \lceil \frac{\bar{c}_u(\mathcal{B})}{\lambda} \rceil$, indicating that c_u reaches or exceeds the λ-multiple successor of $\bar{c}_u(\mathcal{B})$,
  - set $\bar{c}_u(\mathcal{B}) \leftarrow c_u$;
  - for each $q \in \mathcal{B}$, instruct u to send the current counter increment, i.e., $c_u - \bar{c}_u(q)$, to the corresponding coordinator, and set $\bar{c}_u(q) \leftarrow c_u$;
  - for each of those $DT(q)$'s which need to start a new round, invoke the Procedure for New Round Starting;
  - for each of those $DT(q)$'s which become matured, invoke the Procedure for Query Maturity;
  - if the next bucket in $\mathcal{A}(u)$ exists, $\mathcal{B} \leftarrow \mathcal{B}$.next and check $\mathcal{B}$ by repeating this black-point bullet; otherwise, terminate the procedure;
- otherwise (i.e., $c_u < \lambda \cdot \lceil \frac{\bar{c}_u(\mathcal{B})}{\lambda} \rceil$), terminate the procedure.

5.3 Theoretical Analysis

Implementations and Time Complexity. We first analyse the running time cost of each above individual procedure.

Running Time of Query Registrations. For query registrations, the challenge lies in finding the *successor* non-empty bucket $\mathcal{B}$ in $\mathcal{A}(u)$ for a given power-of-two slack value λ, where $\mathcal{B}$ is the *first* non-empty bucket in $\mathcal{A}(u)$ of slack value $\geq \lambda$. Next, we prove the following lemma:

Lemma 5.4. *There exists an implementation which can find the successor bucket in $\mathcal{A}(u)$ for any given power-of-two slack in $O(1)$ amortized expected time with $O(|\mathcal{A}(u)|)$ space.*

Proof. In our implementation, there are two main components: (i) a dynamic universal hashing table on the slack-bucket pairs, each of which corresponds to a bucket in $\mathcal{A}(u)$, and (ii) a bit-array which encodes the emptiness status of all the possible buckets, where the $(i+1)^{\text{st}}$ bit is 1 if and only if the bucket $\mathcal{B}_i$ (of slack 2^i) is non-empty, for all $i = \{0, 1, 2, \ldots, \lfloor \log_2 \tau_{\max} \rfloor\}$.

The Hash Table. With the hash table, in $O(1)$ expected time, we can find the pointer of the bucket (if exists) with respect to a given slack value. Moreover, by standard dynamic universal hashing [7], the hash table can achieve: (i) $O(1)$ expected time for each searching, (ii) $O(1)$ amortized time for each slack-bucket pair insertion and (lazy) deletion, and (iii) $O(|\mathcal{A}(u)|)$ space consumption at all time.

The Bit-Array. Without loss of generality, we assume that $\tau_{\max}$ can be represented with one *word*, because it is fairly easy to extend our technique to the case of $O(1)$-word representable $\tau_{\max}$. As a result, the bit-array encoding the emptiness status of all the possible buckets fits in a one-word integer. Thus, the space consumption of this bit-array is $O(1)$.

Suppose the integer value of the bit-array is x; given a target power-of-two slack $\lambda = 2^i$, the corresponding slack value of the

successor non-empty bucket (if exists) in $\mathcal{A}(u)$ of λ can be found by the following steps:

- let $y \leftarrow x/2^i$, that is to shift all the lower i bits out;
- if $y = 0$, report the successor bucket of λ does not exist in $\mathcal{A}(u)$;
- otherwise, compute $z \leftarrow y \,\&\, \neg(y-1)$;
- return $z \cdot 2^i$ as the slack value of the successor non-empty bucket of $\lambda = 2^i$ in $\mathcal{A}(u)$.

The correctness of the above calculations follows the fact that $\log_2 z$ is the position of the lowest bit of 1 in y. To see this, suppose the lowest bit of 1 is at position j (starting from 0); $y-1$ flips the j^{th} bit from 1 to 0 and all the bits at positions $< j$ from 0 to 1, while all the bits at positions $> j$ remain unchanged. For example, for $y = 10110\underline{1}00$, the lowest bit of 1 is at position $j = 2$, and $y - 1 = 10110\underline{011}$. As a result, $\neg(y-1)$ (e.g., $\neg(y-1) = 01001\underline{100}$) flips all the lowest j bits back to the same as those of the original y, yet all those bits at positions $> j$ are flipped to opposite. Therefore, taking the logic-and operation (&) between y and $\neg(y-1)$ gives a binary string z (e.g., $z = 00000\underline{1}00$) with 1 at position j and all 0's elsewhere. In other words, $\log_2 z = j$, and the correctness the calculations follows. Moreover, all the above calculations can be performed in $O(1)$ time.

Finally, given the slack of the successor bucket $\mathcal{B}$, one can find the pointer of $\mathcal{B}$ with the aforementioned hash table. Putting everything together, the lemma thus follows. □

Thus, by Lemma 5.4, the DT instance for a new query can be constructed in $O(|\mathcal{N}(q)|)$ amortized expected time, linear to the number of participants in this instance.

Running Time of Query Terminations and Maturity. For each query q, by maintaining pointers properly to record the location of query q itself in the bucket $\mathcal{B}$ in each of its canonical node $u \in \mathcal{N}(q)$, locating and removing q from each such bucket can be done in $O(1)$ time. When a bucket becomes empty, the corresponding slack-bucket pair needs to be removed from the hash table, which causes $O(1)$ amortized time. Thus, the running time of each query termination or maturity can be handled in $O(|\mathcal{N}(q)|)$ amortized time.

Running Time of New Round Starting. Consider a query q; when $DT(q)$ needs to start a new round, q needs to be moved from the current bucket to a (possibly new) bucket in each $u \in \mathcal{N}(q)$. As aforementioned, locating q in $\mathcal{B} \in \mathcal{A}(u)$ can be done in $O(1)$ via the pointers. To find the bucket $\mathcal{B}_i$ corresponding to the new slack $\lambda = 2^i$, the procedure scans $\mathcal{A}(u)$ from the current bucket $\mathcal{B}$ towards the head until a "proper position" is found, which is essentially the position of $\mathcal{B}_i$ in (or should be inserted to) the sorted linked list $\mathcal{A}(u)$. Therefore, this cost is bounded by $O(k+1)$, where k is the number of buckets in $\mathcal{A}(u)$ that are "visited" (with slack $\geq \lambda$ and less than the slack of $\mathcal{B}$) and the $O(1)$ term comes from the cost of checking the bucket with slack $< \lambda$ (i.e., the Case 2) and locating q in $\mathcal{B}$.

Let γ be the total number of rounds of $DT(q)$. Summing over all the γ rounds of $DT(q)$, the total cost of starting new rounds within a node $u \in \mathcal{N}(q)$ is bounded by $O\left((\sum_{r=1}^{\gamma} k_r) + \gamma\right)$, where k_r is the number of buckets visited during the scan of $\mathcal{A}(u)$ when starting the r^{th} round. We claim that $O\left((\sum_{j=1}^{\gamma} k_r) + \gamma\right) = O(\log \frac{1}{\varepsilon})$.

Proof of the Claim. By Theorem 5.3, we know that γ is bounded by $O(\log \frac{1}{\varepsilon})$. Next, we bound $\sum_{j=1}^{\gamma} k_r$. First, observe that the buckets visited during the scan of $\mathcal{A}(u)$ must be the buckets corresponding to the slacks falling in the range of $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum possible slacks in $DT(q)$, respectively. Furthermore, according to the P2S-DT algorithm, at the end of each round, the threshold value τ must be decreased by at least a factor of 2, and thus, the slack value in the next round must be at most *half* of the current slack value. In other words, the bucket corresponding to the slack in a new round must be at least "one-step forward" of the bucket $\mathcal{B}$ for the current round in $\mathcal{A}(u)$. As a result, the set of buckets visited during the scan of $\mathcal{A}(u)$ when starting a new different round must be *joint*. Therefore, we have:

$$\sum_{r=1}^{\gamma} k_r \leq \log_2 \lambda_{\max} - \log_2 \lambda_{\min} + 1 \leq \log_2 \lfloor \frac{\tau}{2h} \rfloor_2 - \log_2 \lfloor \frac{\varepsilon \cdot \tau}{2h} \rfloor_2 + 1$$
$$\leq \log_2 \frac{\tau/2h}{\varepsilon \cdot \tau/4h} + 1 = O(\log \frac{1}{\varepsilon}).$$

Note that when a new bucket is created, a new slack-bucket pair is inserted to the aforementioned universal hash table, whose cost is bounded by $O(1)$ amortized. Therefore, by the above claim, the overall maintenance cost of new round starting's in $DT(q)$ is thus bounded by $O(|\mathcal{N}(q)| \cdot \log \frac{1}{\varepsilon})$ amortized.

Running Time of Counter Increments. Next, we analyse the running time bound for handling each counter increment. First, observe that the cost of instructing the queries in the buckets to notify their coordinators can be charged to the "communication cost" of each of the corresponding DT instances. Second, as the costs of handling new round starting's and maturity have been bounded separately, it suffices to bound the total cost of the scan of $\mathcal{A}(u)$. Clearly, the number of buckets that have been visited during the scan of $\mathcal{A}(u)$ for handling a counter increment on c_u is $k_u + 1$, where k_u is the number of buckets that have at least one query to notify the corresponding coordinator, and the "1" is for the first bucket which does not satisfy the notification condition. As a result, the cost of $O(k_u)$ can be charged to the communication cost of those DT instances in the buckets, and the $O(1)$ cost can be charged to the counter increment, which in turn can be charged to the cost of element processing.

According to Theorem 5.3, the communication cost of $DT(q)$ is bounded by $O(|\mathcal{N}(q)| \cdot \log \frac{1}{\varepsilon})$. By Fact 3, we have $|\mathcal{N}(q)| = O(\log^d N)$ and the total processing cost for n elements is bounded by $O(n \cdot \log^d N)$. Combining the costs of the procedures for query maturity and new round starting, for m queries and n elements, the overall cost of the procedure for counter increments is bounded by $O(m \cdot \log^d N \cdot \log \frac{1}{\varepsilon} + n \cdot \log^d N)$.

Space Consumption. Recall that for each last-layer node u, the data structures for the bucketing technique only take $O(|\mathcal{A}(u)|)$ space. Summing over all such nodes gives the overall space consumption of the DT Manager module bounded by $O(m_{\text{alive}} \cdot \log^d N)$, where m_{alive} is the number of the alive queries.

The Overall Cost of the DT Manager Module. We summarize the key results of the DT Manager module in the theorem below:

Theorem 5.5. *For m queries (along with their registrations and terminations) and n elements, the DT Manager module achieves:*

- $O(m \cdot \log^d N \cdot \log \frac{1}{\varepsilon} + n \cdot \log^d N)$ *time in expectation*[2]*, and*
- $O(m_{alive} \cdot \log^d N)$ *space consumption at all time.*

Remark. We note that our bucketing technique will be of independent interests, as it is applicable not only to our *FastRTS* , but also to all those algorithms which organize DT instances with heaps, such as the *QGT* algorithm and a very recent work [23]. In this case, the bucketing technique can remove the logarithmic factor caused by the use of the heaps.

6 EFFECTIVE OPTIMIZATIONS

In this section, we introduce two powerful and effective optimizations for *FastRTS* , which substantially reduce our algorithm's actual running time and the peak memory usage over the entire process. Meanwhile, all the theoretical bounds of *FastRTS* retain.

6.1 The Range Shrinking Technique

In this subsection, we introduce a technique, called *Range Shrinking*, for reducing the number of participants in $DT(q)$ for query q.

The Technique. Consider a query q with range $R(q)$ and threshold $\tau(q)$. The basic idea is to first *extend* q to a *super query* $\tilde{q}$ whose range is a super range of of $R(q)$, namely, $R(q) \subseteq R(\tilde{q})$, and run the DT instance with $\tilde{q}$. The rationale here is that q is ε-matured *only if* the total weight of elements falling in $R(\tilde{q})$ is at least $(1-\varepsilon)\tau(q)$ since q was registered. As a result, instead of running $DT(q)$ directly, we can first run a DT instance for a super query $\tilde{q}$ with $\mathcal{N}(\tilde{q})$ as participants and with threshold $\tau(\tilde{q}) = \tau(q)$. Based on this observation, indeed, we can run DT instances with respect to a sequence of super queries of q before actually running $DT(q)$. The detailed steps of our Range Shrinking technique are as follows.

- Let $\bar{s}$ be the precise counter sum of all the nodes $u \in \mathcal{N}(q)$ in the *IncSegTree* at the moment when q is registered.
- Initialize a super query $\tilde{q}$ of q with range $R(\tilde{q}) \supseteq R(q)$ and threshold $\tau(\tilde{q}) = \tau(q)$.
- Repeat the following loop for at most $\lfloor \log_2 \frac{1}{\varepsilon} \rfloor$ times:
  - run $DT(\tilde{q})$ to capture an ε'-maturity with $\varepsilon' = \varepsilon\tau(q)/\tau(\tilde{q})$;
  - collect the precise counter sum s of all the nodes $u \in \mathcal{N}(q)$;
  - if $\tau(q) - s + \bar{s} \le \varepsilon\tau(q)$, report the ε-maturity of q and return;
  - otherwise,
    - *shrink* $R(\tilde{q})$ to a smaller range R such that $R(q) \subseteq R \subseteq R(\tilde{q})$;
    - update $\tilde{q}$ by setting $R(\tilde{q}) \leftarrow R$ and $\tau(\tilde{q}) \leftarrow \tau(q) - s + \bar{s}$;
    - repeat the loop;
- If the procedure has not stopped yet,
  - collect the precise counter sum s of all the nodes $u \in \mathcal{N}(q)$;
  - update $\tilde{q}$ by setting $R(\tilde{q}) \leftarrow R(q)$ and $\tau(\tilde{q}) \leftarrow \tau(q) - s + \bar{s}$;
  - run $DT(\tilde{q})$ to capture an ε'-maturity with $\varepsilon' = \varepsilon\tau(q)/\tau(\tilde{q})$;
  - report the ε-maturity of q.

Correctness. The correctness of the Range Shrinking technique follows from two facts. First, the total counter increment in $DT(\tilde{q})$ is at most $\tau(q) - s + \bar{s}$, i.e., the gap between the threshold $\tau(q)$ and the total counter increment happened in $R(q)$ so far. Second, $R(\tilde{q}) \supseteq R(q)$ implies the total counter increment in the former must be at least that of the latter. Thus, the ε-maturity period of $DT(q)$ will not be missed when $DT(\tilde{q})$ is running.

[2]The expectation restriction comes from the possible hash table lookups for query registrations. Except this, the overall time bound is for the worst case.

Furthermore, in the loop, every time the ε'-maturity of $DT(\tilde{q})$ is reported, a safety check on whether $\tau(q) - s + \bar{s} \le \varepsilon\tau(q)$ is performed. This ensures the ε-maturity of q can be reported correctly in the loop. Finally, the last $DT(\tilde{q})$ outside the loop essentially tracks the remaining gap between $\tau(q)$ and the total counter increment in $R(q)$ so far. An ε-maturity of $DT(q)$ must also be correctly captured in this case.

Running Time Analysis. Denote the super query in the i^{th} loop by $\tilde{q}_i$, and the last super query outside the loop by $\tilde{q}_{\text{last}}$. Observe that the number of rounds in each $DT(\tilde{q}_i)$ is at most $O(\log \frac{1}{\varepsilon})$, because $\tau(\tilde{q}_i) \le \tau(q)$ always holds, and thus, $\varepsilon' \le \varepsilon$ holds. Likewise, $DT(\tilde{q}_{\text{last}})$ also has at most $O(\log \frac{1}{\varepsilon})$ rounds. In addition, since $R(\tilde{q}_{\text{last}}) = R(q)$, we have $|\mathcal{N}(\tilde{q}_{\text{last}})| = |\mathcal{N}(q)|$. Therefore, the total running time cost of the DT instances of all these super queries is bounded by $O\left((\sum_i |\mathcal{N}(\tilde{q}_i)| + |\mathcal{N}(q)|) \cdot \log \frac{1}{\varepsilon}\right)$. Furthermore, each collection of the precise counter sum s for the query q takes $O(|\mathcal{N}(q)|)$ time, and there are at most $O(\log \frac{1}{\varepsilon})$ such collections.

Adding up these two costs, the running time of the entire Range Shrinking process is bounded by $O\left((\sum_i |\mathcal{N}(\tilde{q}_i)| + |\mathcal{N}(q)|) \cdot \log \frac{1}{\varepsilon}\right)$.

Next, we show that, by a careful strategy for constructing the super queries, $\sum_i |\mathcal{N}(\tilde{q}_i)|$ can be bounded by $O(\log^d N)$. Therefore, the above overall cost remains $O(\log^d N \cdot \log \frac{1}{\varepsilon})$ as before.

Constructing the Super Queries. We first revisit the algorithm to obtain the canonical node set $\mathcal{N}(q)$ for a query q with a Segment Tree. This process is essentially a traversal on the tree:

- let u denote the node that is currently visited; initially, u is the root of the base tree on the current dimension j ($j \in \{1, 2, \ldots, d\}$);
- define $R_j(u)$ and $R_j(q)$, respectively, to be the projections of the ranges $R(u)$ and $R(q)$ on the j^{th} dimension;
- if $R_j(u) \subseteq R_j(q)$,
  - if this is at the last layer, i.e., $j = d$, add u to $\mathcal{N}(q)$;
  - otherwise, recursively traverse the associated base tree on the next dimension of u;
- if $R_j(u) \cap R_j(q) \neq \emptyset$, continue the traversal to both u's child nodes if they exist, and check these child nodes as the current node accordingly;
- otherwise, i.e., $R_j(u) \cap R_j(q) = \emptyset$, stop the traversal at u.

It is known that for a base tree on dimension j, at each level, there can be at most two nodes with ranges fully contained in $R_j(q)$. Thus, at most $O(h)$ nodes would recurse into the next dimension until to the last dimension (i.e., the last layer), where $O(h)$ is an upper bound on the height of all the base trees. As a result, $|\mathcal{N}(q)|$ is bounded by $O(h^d)$. In particular, in an *IncSegTree*, $O(h) = O(\log N)$.

To construct a super query $\tilde{q}_i$, we modify the above traversal algorithm slightly. We *conceptually* "truncate" all the base trees in *IncSegTree* at the depth of 2^i for $i \in \{1, 2, \ldots, \lfloor \log \frac{1}{\varepsilon} \rfloor\}$, i.e., the height of every base tree becomes 2^i. We traverse this "truncated" *IncSegTree* as if we compute the canonical node set for q.

Specifically, consider a currently visited "leaf" node u (at level-2^i) in a base tree on dimension j; if u satisfies $R_j(u) \cap R_j(q) \neq \emptyset$, recurse into the base tree on next dimension if $j < d$ or add u to $\mathcal{N}(\tilde{q}_i)$ if this is the last layer, i.e., $j = d$. Finally, we set $R(\tilde{q}_i) = \cup_{u \in \mathcal{N}(\tilde{q}_i)} R(u)$.

Example 3. *Consider the example shown in Figure 1; to construct the super query $\tilde{q}_1$ of q with respect to the depth of 2^i, the traversal*

on the base tree on the first dimension in Figure 1(a) stops at the two nodes with ranges $[1, 9)$ *and* $[9, 16]$. *Moreover, as for the second dimension, the traversal stops at the node* $[1, 9)$ *in Figure 1. Therefore, we construct* $R(\tilde{q}_1) = [1, 9) \times [1, 9) \cup [9, 16] \times [1, 9) = [1, 16] \times [1, 9)$ *as shown in Figure 1(c). As for the next shrinking, since the truncation depth is at* $2^2 = 4 = \log_2 N$, $\tilde{q}_2$ *degenerates back to* q.

As per the above construction, three properties hold for all $\tilde{q}_i$: (i) $|\mathcal{N}(\tilde{q}_i)| = O\left((2^i)^d\right)$, (ii) $|\mathcal{N}(\tilde{q}_i)| \leq |\mathcal{N}(q)|$, and (iii) $R(\tilde{q}_i) \supseteq R(q)$. As a result, by (iii), each $\tilde{q}_i$ is a valid super query of q. By (i) and (ii), we have $\sum_i |\mathcal{N}(\tilde{q})_i| = O(\sum_i (2^i)^d) = O(|\mathcal{N}(q)|) = O(\log^d N)$. Therefore, the overall cost of running $DT(q)$ with the Range Shrinking technique is still bounded by $O(\log^d N \cdot \log \frac{1}{\varepsilon})$.

Space Consumption. Since the Range Shrinking process needs to collect the precise counter sum for the query q, the *IncSegTree* still has to materialize $O(|\mathcal{N}(q)|)$ nodes for q to support this operation. Also, the *IncSegTree* needs to materialize $O(|\mathcal{N}(\tilde{q}_i)|)$ nodes for each super query $\tilde{q}_i$. Nonetheless, as per the analysis of the running time, the total number of all these nodes is at most $O(\log^d N)$. The overall space consumption $O(m_{\text{alive}} \cdot \log^d N)$ bound does not change.

Benefits. First, in terms of running time, most counter increments out of $\tau(q)$ are tracked with a considerably smaller DT instance $DT(\tilde{q}_i)$. Moreover, it also provides a chance of "early termination", in the sense that an ε-maturity of q can be captured even without actually running $DT(q)$. As a result, the actual running time is substantially reduced. Second, in terms of space consumption, it consists of two main parts: (i) the space of the base tree of *IncSegTree*, and (ii) the space of all the DT instances. As for the base tree, while in theory *IncSegTree* needs to materialize $O(m \cdot \log^d N)$ nodes for m queries, in practice these nodes are largely shared such that the actual number of the nodes materialized is often much smaller than this worst-case bound. In contrast, the space bound on the total size of all the DT instances, i.e., $O(m \log^d N)$, is pretty solid, as the sizes cannot be shared: one has to store q in each of the nodes in $\mathcal{N}(q)$. Thus, the space consumption is dominated by the total size of the DT instances. Our Range Shrinking technique can keep each DT instance small most of the time. More importantly, it makes the *peak memory usage* of each DT instance *asynchronous*, in the sense that, the DT instances seldom have their maximum sizes, i.e., running with $|\mathcal{N}(q)|$ participants, at the same time. Our experimental results show that, *the Range Shrinking technique can effectively reduces both the actual running time and space consumption of* FastRTS .

6.2 Further Reducing the Memory Footprint

As aforementioned, the Range Shrinking technique allows *FastRTS* to run DT with super queries and thus, reduce the actual space consumption by asynchronizing the peak memory usage of the DT instances. However, the *IncSegTree* still has to materialize $O(|\mathcal{N}(q)|)$ nodes for each query q so as to support the collection of precise counter increment for the range $R(q)$. In other words, even though an ε-maturity of q can be captured by the DT instance of some super query, the nodes of $\mathcal{N}(q)$ still have to be materialized. Therefore, the overall *memory footprint* (i.e., the peak memory usage) may not be desired, especially when the number m of queries is large. Next, we propose an optimization, called the *Range Counting* technique, to address this issue.

The basic idea of this optimization is to make the *IncSegTree* *"purely incremental"*, which means the *IncSegTree* does not need to materialize those $O(|\mathcal{N}(q)|)$ nodes for q at the moment when q is registered. Instead, it just materializes the nodes for the super query $\tilde{q}_i$ whose DT instance is currently running for q and those nodes for q's super queries $\tilde{q}_j$ for all $j < i$. However, the challenge is to support finding precise counter increments for ranges.

Our solution to this challenge is to maintain a standard *Range Tree* [13] on the stream elements to support *range counting*'s, which report the total weight sum of all the points (in the tree) falling into the given range. Roughly speaking, this Range Tree $\mathcal{T}$ is constructed from an empty tree. For each stream element e arrives, insert the point $v(e)$ with weight $w(e)$ into $\mathcal{T}$. When $\mathcal{T}$ has more than $c \cdot m_{\text{alive}}$ (for some constant $c > 0$) points, destroy $\mathcal{T}$ (all the current points are discarded) and maintain $\mathcal{T}$ from an empty tree for the *new* elements from the stream. The concrete algorithm works as follows.

- For each query q, maintain two information:
  - $\bar{s}(q)$: the weight sum of the points falling in $R(q)$ in the current Range Tree at the moment when q is registered; if the current Range Tree is constructed (from an empty tree) after q's registration, $\bar{s}(q) = 0$; and
  - $s_{\text{prv}}(q)$: the total counter increments for $R(q)$ in the *previous* Range Trees which have been destroyed after q's registration; if there is no such Range Tree, $s_{\text{prv}}(q) = 0$.

  Therefore, the total counter increments happened in $R(q)$ so far can be calculated by $s(q) - \bar{s}(q) + s_{\text{prv}}(q)$, where $s(q)$ is the weight sum of the points in the current Range Tree $\mathcal{T}$ falling in $R(q)$.
- When a query q is registered,
  - perform a range counting for $R(q)$ in $\mathcal{T}$ to obtain $\bar{s}(q)$, and initialize $s_{\text{prv}}(q) \leftarrow 0$;
  - run $DT(q)$ with the Range Shrinking technique, yet the *IncSegTree* is now purely incremental;
- When a stream element e arrives,
  - process e with the *IncSegTree* as before;
  - insert $v(e)$ with weight $w(e)$ to the current Range Tree $\mathcal{T}$;
  - if $\mathcal{T}$ contains more than $c \cdot m_{\text{alive}}$ points,
    - for each *alive* query q, perform a range counting for $R(q)$ to obtain $s(q)$ in $\mathcal{T}$; update $s_{\text{prv}}(q) \leftarrow s(q) - \bar{s}(q) + s_{\text{prv}}(q)$, and $\bar{s}(q) \leftarrow 0$;
    - destroy $\mathcal{T}$ and set $\mathcal{T} \leftarrow \emptyset$;

Running Time Analysis. First, observe that when the current Range Tree $\mathcal{T}$ is destroyed, the cost of the $O(m_{\text{alive}})$ range counting operations for the alive queries can be charged to the cost of the insertions of the $\Omega(m_{\text{alive}})$ points in $\mathcal{T}$. By the standard results [13], each range counting and each point insertion can be performed in $O(\log^d m_{\text{alive}})$ time in a Range Tree with at most $O(m_{\text{alive}})$ points. As a result, each element e can be processed in the Range Tree $\mathcal{T}$ in $O(\log^d m_{\text{alive}}) = O(\log^d N)$ amortized time. Moreover, by the fact that the processing cost for each element in the *IncSegTree* is bounded by $O(\log^d N)$, the maintenance of the Range Tree does not affect the overall running time bound as before.

Space Consumption. By the standard results [13], it is known that the space consumption of the Range Tree is bounded by $O(m_{\text{alive}} \cdot \log^{d-1} m_{\text{alive}})$. Furthermore, as the *IncSegTree* is now purely incremental, the overall space consumption bound can be written

as $O(\sum_{\text{alive query } q} |\mathcal{N}(\tilde{q})| + m_{\text{alive}} \cdot \log^{d-1} m_{\text{alive}})$. The worst-case bound $O(m_{\text{alive}} \cdot \log^d N)$ still holds.

Benefits. By the above Range Counting technique, the space consumption now is just the total size of all the DT instances for the super queries which are run so far plus the space of the Range Tree. Therefore, this space consumption can be substantially smaller than the space of the previous *IncSegTree*. As shown in our ablation study in the experiment section, this technique can *significantly improve the performance of* FastRTS *by orders of magnitude in terms of both overall running time and the memory footprint.*

7 EXPERIMENTS

The performance evaluation of our *FastRTS* consists of three parts. First, we conduct comprehensive experiments on synthetic datasets with dimensionality $d = 1, 2, 3, 4$. Second, we run experiments on real stock trading data. Last but not least, we perform an ablation study on our proposed optimization techniques.

Competitors. We compare the following methods.

- ***FastRTS*** : Our *FastRTS* equipped with both the Range Shrinking and Range Counting techniques.
- ***QGT*** : the state-of-the-art *QGT* algorithm.
- ***SegInv***: a conventional stabbing-based approach with a Segment-Interval tree [13] which is a Segment Tree with an Interval Tree as the base tree on the last dimension. The space consumption of *SegInv* is $O(m \cdot \log^{d-1} m)$ and the overall running time is $O(m \cdot \tau_{\max} + n \cdot \log^d m)$.
- ***Rtree***: another conventional stabbing-based approach with an R-tree [5, 17]. The space consumption of this method is $O(m)$, while the worst-case time complexity is $O(n \cdot m + m \cdot \tau_{\max})$.

All the above algorithms are implemented in C++ and complied by gcc 9.3.0 with flag O3 turned on. The source code are at here [1].

Machine and OS. All of the experiments were conducted on a machine equipped with an Intel Xeon(R) W-2145 CPU @ 3.70GHz and 100GB RAM running Ubuntu 20.04.3.

Evaluation Metric. We evaluate the performance of an algorithm by measuring the *overall running time* and the *peak memory usage* (i.e. the memory footprint). In all the following experiments, a competitor algorithm may not have experimental results in certain diagram with tasks under certain parameter settings. This is caused by either the algorithm fails to complete the task within 10 hours (i.e., 36000 seconds), or its peak memory usage exceeds 100GB breaking down our machine.

Data Generation. In all the experiments, data are generated as follows, unless specified otherwise in the experiment setup.

Element Generation. For each element e, the value $v(e)$ is a point uniformly at random picked in $\mathbb{U}^d$, and the weight $w(e) = \lfloor x + 0.5 \rfloor$, where x is sampled from the Gaussian distribution with $\mu = 10$ and $\sigma = 1$. If $w(e) \notin \mathbb{U}$, generate $w(e)$ again.

Query Generation. All queries have a same threshold $\tau = m$. The range $R(q)$ of each query q is a hypercube (rsp., an interval when $d = 1$ and a square when $d = 2$) in $\mathbb{U}^d$, whose volume is 1% of the entire data space. That is, the side length of $R(q)$ is $\ell = 0.01^{\frac{1}{d}} N$. Given a d-dimensional point $\vec{z}$ as a *seed*, the center of the range $R(q)$ is generated by a d-dimensional Gaussian distribution with *mean* $\vec{\mu} = \vec{z}$, where all the dimensions are independent and each of them has a same standard variance $\sigma = \frac{0.15\ell}{\sqrt{d}}$. In the experiments for synthetic datasets and ablation study, we set $\vec{z}$ to be the center of the entire space, that is $\vec{z}$ has coordinates all equal to $N/2$. Note that we will have different setting for the seed $\vec{z}$ in the experiments on real data. If a generated range $R(q)$ is not fully contained in $\mathbb{U}^d$, then re-generate $R(q)$ by the same process.

Query Registrations and Terminations. As per the above query generation process, each query q is stabbed by an element e with probability of 1%, and the expectation of $w(e)$ is 10. As a result, in expectation, q will be matured after $\frac{\tau}{1\% \cdot 10} = 10\tau$ elements since its registration. However, at each of these 10τ time stamps, we set q to have probability p to be terminated before its maturity. This probability p is set properly such that the probability that q remains alive for 10τ time stamps without being terminated is 20%. As for query registrations, all the experiments are with a *hot-start* setup, that is, there are m queries registered at time stamp 0 at the beginning. Since then, another m queries will be registered within the first 10τ time stamps. To achieve this, we registered these m queries following a *Poisson Process* with an *intensity* of $\frac{m}{10\tau}$. As a result, over the entire process, there will be in total $2m$ queries registered. To ensure the later registered queries have enough chance to get matured, we set the stream length as $n = 20\tau$.

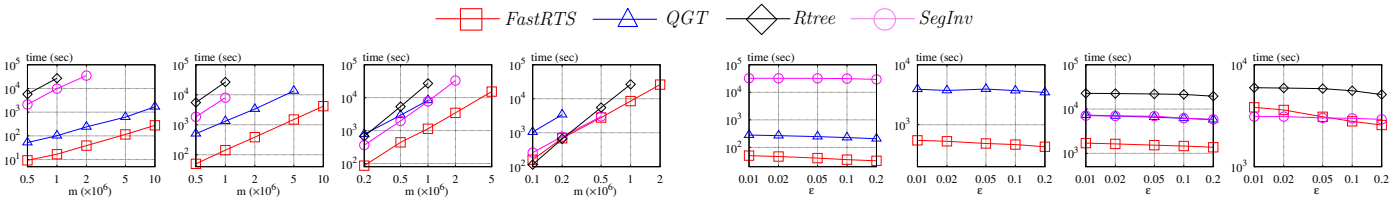
7.1 Evaluation on Synthetic Data

Parameter Settings. We conduct experiments on synthetic datasets with the following parameter settings, where K= 10^3 and M= 10^6.

- $N = |\mathbb{U}| = 10\text{M}$, $\tau = m$, $n = 20\tau$, $d \in \{1, 2, 3, 4\}$
- for $d = 1$ or 2, $m \in \{500\text{K}, 1\text{M}, \mathbf{2M}, 5\text{M}, 10\text{M}\}$
- for $d = 3$, $m \in \{200\text{K}, 500\text{K}, \mathbf{1M}, 2\text{M}, 5\text{M}\}$
- for $d = 4$, $m \in \{100\text{K}, 200\text{K}, \mathbf{500K}, 1\text{M}, 2\text{M}\}$
- $\varepsilon \in \{0.01, 0.02, \mathbf{0.05}, 0.1, 0.2\}$

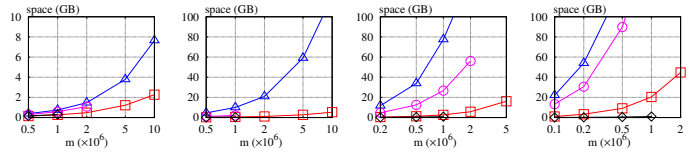
The default value of each parameter is highlighted in bold. When varying a parameter, all the others are set to their default values.

Comparisons on Overall Running Time. Figures 2 (a) - (d) show the overall running time of the competitors when varying m on different dimensionality d. From these figures, we have the following observations. First, our *FastRTS* is the only one that can complete all the experiment tasks. Second, over all the four dimensionalities, *FastRTS* consistently outperforms *QGT* by around an order of magnitude. Third, *FastRTS* outperforms the two conventional stabbing-based approaches by up to *three* (for $d = 1$) and *two* for ($d = 2$) orders of magnitude. In particular, while *FastRTS* can complete the running tasks with the default m within 100 seconds for $d = 1$ (rsp., 1000 seconds for $d = 2$), these two competitors require about 10 hours for *SegInv* and even more for *Rtree*. This nicely illustrates the superiority of our theoretical running time bound. Furthermore, *QGT* is also up to 10 times faster than these two methods. However, when $d \geq 3$, *Rtree* and *SegInv* start becoming competitive to *FastRTS* . In particular, *FastRTS* even slightly losts to *SegInv* when $d = 4$. One possible reason is that the dataset sizes are considerably small when $d \geq 3$, as we hope to have more competitors completing the running tasks. Another possible reason



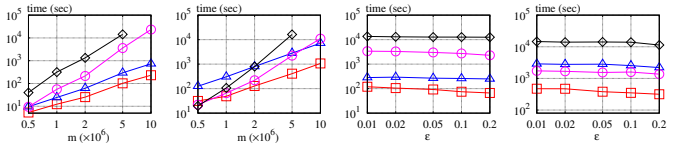
(a) $d = 1$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$ (e) $d = 1$ (f) $d = 2$ (g) $d = 3$ (h) $d = 4$

Figure 2: Overall running time v.s. m [(a) - (d)] and v.s. ε [(e) - (h)] on synthetic datasets



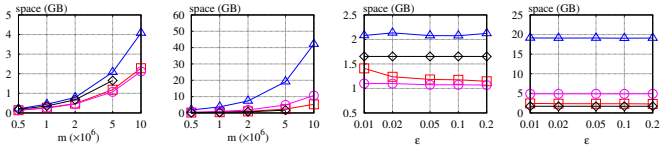
(a) $d = 1$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$

Figure 3: Peak memory usage v.s. m on synthetic datasets



(a) $d = 1$ (b) $d = 2$ (c) $d = 1$ (d) $d = 2$

Figure 4: Running time v.s. m and v.s. ε on real datasets



(a) $d = 1$ (b) $d = 2$ (c) $d = 1$ (d) $d = 2$

Figure 5: Peak memory usage v.s. m and v.s. ε on real datasets

is that with d increasing, the poly-logarithmic gets worse exponentially, which severely impacts the performance of both *FastRTS* and *QGT* algorithms.

Furthermore, Figure 2(e) - (h) show the running time when varying ε. From these figures, *FastRTS* clearly outperforms all the other competitors who completed the running tasks when $d \leq 3$. But *FastRTS* is inferior to *SegInv* when $d = 4$. Moreover, as expected, all the four competitors run faster when ε increases.

Comparisons on Peak Memory Usage. Next, we look into the space consumption. As shown in Figure 3, *Rtree* is the most space-efficient. This is obvious because the space complexity of *Rtree* is just $O(m)$. However, as aforementioned, *Rtree* is also the slowest among the four in general. On the other hand, the peak memory usages of both *SegInv* and *QGT* increase quickly with the dimensionality d increasing; both of them quickly use up the 100GB memory on relatively small datasets: $d = 3$ with $m = 2$M and $d = 4$ with $m = 0.5$M for *QGT* and $m = 1$M for *SegInv*. In contrast, the peak memory usages of our *FastRTS* on these settings are much more friendly; they are: 5.77GB, 9.11GB and 20GB, respectively. The memory usage trend of *FastRTS* increases much slower than these two algorithms. From Figure 3, on some settings, *FastRTS* outperforms these two algorithms by up to two orders of magnitude.

In summary, in these synthetic experiments, *FastRTS* is much more scalable and stable over all various settings than the others competitors, in terms of both running time and space consumption.

7.2 Evaluations on Real Stock Trading Data

Next, we investigate the performance of *FastRTS* on real datasets.

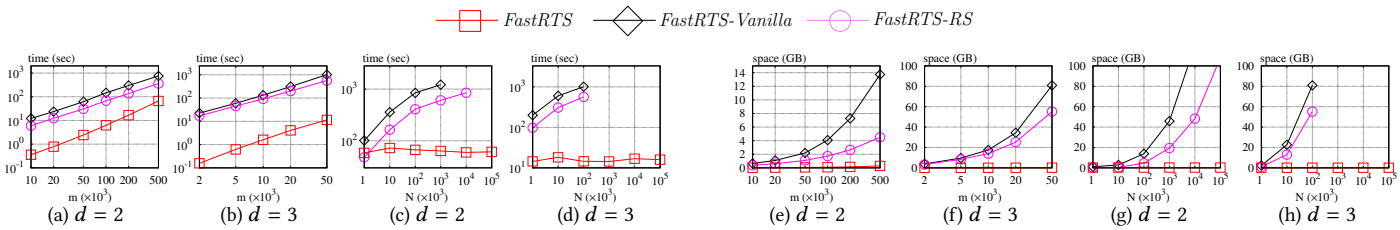
Data Description. We run experiments on the real trading history of two Stocks, A and B. These data [1] are the transactions of Stock A (rsp. Stock B) from 2015 to 2021. For each transaction of A, we extract a stream element e with $v(e)$ equal to the *price* information and $w(e)$ equal to the *volume* (i.e., number of shares) information. Thus, a 1-dimensional stream w.r.t. the transaction history of A is generated. To generate a 2-dimensional stream, for each element e generated for the transaction of A, we search the *predecessor* transaction of B in terms of time stamp, and extract the price t in this transaction of B as the second dimension value of e. Thus, a 2-dimensional element e' with value $v(e') = (v(e), t)$ and weight $w(e') = w(e)$ is generated. In this way, we simulate the 2-dimensional Range Thresholding query example mentioned in the Introduction.

Query Generation. The queries in this set of experiments are generated in the same way as before. Here, the seed $\vec{z}$ set as the *average* of all the element points for the first m queries which will be registered at the beginning; $\vec{z}$ is set as the i^{th} element point if a query is generated to be registered at time stamp i.

Parameter Settings (on the real stock trading data).

- $N = |\mathbb{U}| = 100$K, $\tau = m$, $n = 20\tau$
- $d \in \{1, 2\}$
- $m \in \{500\text{K}, 1\text{M}, 2\text{M}, \mathbf{5M}, 10\text{M}\}$
- $\varepsilon \in \{0.01, 0.02, \mathbf{0.05}, 0.1, 0.2\}$

Overall Running Time. Figures 4 (a) and (b) show the results of the overall running time v.s. m, the number of queries on the real datasets. As expected, *FastRTS* is consistently the fastest approach on both $d = 1$ and $d = 2$. Interestingly, while *QGT* is the second best performer on $d = 1$, it is outperformed by *SegInv* on $d = 2$. At first glance, this looks inconsistent with the comparisons on the synthetic datasets. Looking into this case, we find this is caused by the *element weights*. Although the query thresholds τ's in both experiments have similar values, the average weight of the elements are quite different. Specifically, the average element weight of the real data is about 400; the expected weight of the elements in synthetic datasets is 10. As a result, although the two τ have the same value, the *effective* threshold in the real data is much smaller. Recall that the time complexity of *SegInv* is often dominated by $O(m \cdot \tau)$, so with an effectively much smaller τ, *SegInv*'s performance gets



(a) $d = 2$ (b) $d = 3$ (c) $d = 2$ (d) $d = 3$ (e) $d = 2$ (f) $d = 3$ (g) $d = 2$ (h) $d = 3$

Figure 6: Ablation Study on running time v.s. m and N [(a) - (d)], and on space v.s. m and N [(e) - (h)]

better. Furthermore, from Figures 4 (c) and (d), all the algorithms run faster with larger ε as expected.

Peak Memory Usage. As shown in Figure 5, the peak memory usage comparisons on the real datasets are consistent with those on the synthetic ones on $d = 1$ and $d = 2$ (shown in Figure 3 (a) and (b)). Specifically, the space consumption of the *QGT* algorithm is consistently the largest with a clear gap from those of the other three competitors. Observe that the space consumption of *SegInv* is smaller than that of *Rtree* on $d = 1$, since both the space consumptions of them are linear to m in this case. However, when $d = 2$, this ranking reverses, as the space consumption bound of *Rtree* is not affected by d but that of *SegInv* grows exponentially fast with d. Interestingly, despite of the theoretical space bound which also grows exponentially with d, the peak memory usage of our *FastRTS* is consistently between those of *SegInv* and *Rtree*, and very close to the winner between the two in both the cases of $d = 1$ and $d = 2$. This is because our optimization techniques enable *FastRTS* to avoid the worst case, and hence, perform much better in practice than as the theoretical bound suggests. This clearly shows the effectiveness of our optimization techniques.

7.3 Ablation Study

The last set of experiments is an ablation study on the effectiveness of our optimization techniques. The competitors in these experiments are the variants of *FastRTS* only. In addition to the "fully-geared" *FastRTS* , we also consider (i) *FastRTS-RS* that is equipped with the Range Shrinking technique only, and (ii) *FastRTS-Vanilla* that is the *raw* version of the algorithm.

Moreover, it is worth mentioning that in these experiments, we intended to use small m to minimize its impact to the actual performance such that we could have a clearer view on the impact of the universe size N. Given that, the parameter setting is as follows:

- $\tau = m$, $n = 20\tau$, $\varepsilon = 0.05$
- $N \in$ {1K, 10K, **100K**, 1M, 10M, 100M}
- for $d = 2$, $m \in$ {10K, 20K, 50K, 100K, 200K, **500K**}
- for $d = 3$, $m \in$ {2K, 5K, 10K, 20K, **50K**, 100K}

Overall Running Time. From Figures 6(a) - (d), with no surprise, the fully-geared *FastRTS* consistently outperforms the other two versions by up to two orders of magnitude, in terms of efficiency. This clearly shows the significance of the two optimization techniques on improving the efficiency. Moreover, *FastRTS-RS* is faster than the raw *FastRTS-Vanilla* on all settings. However, the gap is just not as significant. A possible explanation is that although the Range Shrinking technique allows *FastRTS-RS* to track a considerable portion of counter increments out of $\tau(q)$ with relatively smaller DT instances, materializing all the nodes in $\mathcal{N}(q)$ is still a considerable burden on the overall running time. On the other hand, this is actually a strong evidence of the effectiveness of Range Counting technique. Apart from this, from Figures 6(c) and (d), we can see that the running time of *FastRTS* is not affected by the universe size N as much as suggested by the theoretical bound. Once again, another evidence of the effectiveness of our optimizations.

Peak Memory Usage. As shown in Figures 6(e) - (h), the fully-geared one is always far at the bottom with a dramatic gap from the other two versions. The capability of avoiding materializing the nodes in $\mathcal{N}(q)$ of our Range Counting technique is just powerful. While *FastRTS-RS* is also largely better than the vanilla version on space consumption, for $N = 100$M, it still has to run out of the 100GB memory even just on $m = 50$K when $d = 3$.

From all the above comparisons in the ablation study, the fully-geared *FastRTS* is way better than the other, showing that our optimizations are extremely effective.

8 CONCLUSION

In this paper, we propose a new algorithm, called *FastRTS* , for solving the approximate Range Thresholding (RT) problem. In theory, our *FastRTS* improves the state-of-the-art algorithm by reducing the exponential dependence on the data dimensionality d for the logarithmic factor, yet with a sacrifice of slightly increasing the logarithmic term. The crucial idea to make this improvement happen is our *bucketing technique*. This technique can remove the logarithmic overhead caused by the use of heaps. We note that our bucketing technique is of independent interests as it is also applicable to eliminate a logarithmic overhead for all those algorithms in similar scenarios. In practice, we proposed two extremely effective optimizations to significantly improve the performance of *FastRTS* . We conduct extensive experiments on both synthetic and real datasets. Experimental results show that *FastRTS* outperforms the state-of-the-art competitors by orders or magnitude in terms of both overall running time and space consumption.

REFERENCES

[1] [n.d.]. FastRTS source code and experiment dataset. https://anonymous.4open.science/r/FastRTS.

[2] Daniel J. Abadi, Donald Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stanley B. Zdonik. 2003. Aurora: a new model and architecture for data stream management. *The VLDB Journal* 12, 2 (2003), 120–139.

[3] Arvind Arasu and Jennifer Widom. 2004. A Denotational Semantics for Continuous Queries over Streams and Relations. *SIGMOD Record* 33, 3 (2004), 6–12.

[4] Lars Arge and Jan Vahrenhold. 2004. I/O-efficient dynamic planar point location. *Comput. Geom.* 29, 2 (2004), 147–162. https://doi.org/10.1016/j.comgeo.2003.04.001

[5] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. 1990. The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. In *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, USA, May 23-25, 1990*, Hector Garcia-Molina and H. V. Jagadish (Eds.). ACM Press, 322–331. https://doi.org/10.1145/93597.98741

[6] Jon Louis Bentley and James B. Saxe. 1980. Decomposable Searching Problems I: Static-to-Dynamic Transformation. *J. Algorithms* 1, 4 (1980), 301–358. https://doi.org/10.1016/0196-6774(80)90015-2

[7] Larry Carter and Mark N. Wegman. 1979. Universal Classes of Hash Functions. *J. Comput. Syst. Sci.* 18, 2 (1979), 143–154. https://doi.org/10.1016/0022-0000(79)90044-8

[8] Jianjun Chen, David J. DeWitt, Feng Tian, and Yuan Wang. 2000. NiagaraCQ: A Scalable Continuous Query System for Internet Databases. In *Proceedings of ACM Management of Data (SIGMOD)*. 379–390.

[9] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms, 3rd Edition.* MIT Press. http://mitpress.mit.edu/books/introduction-algorithms

[10] Graham Cormode, S. Muthukrishnan, and Ke Yi. 2011. Algorithms for distributed functional monitoring. *ACM Trans. Algorithms* 7, 2 (2011), 21:1–21:20. https://doi.org/10.1145/1921659.1921667

[11] Graham Cormode, S. Muthukrishnan, and Ke Yi. 2011. Algorithms for Distributed Functional Monitoring. *ACM Trans. Algorithms* 7, 2, Article 21 (March 2011), 21:1–21:20 pages.

[12] Gianpaolo Cugola and Alessandro Margara. 2012. Processing flows of information: From data stream to complex event processing. *Comput. Surveys* 44, 3 (2012), 15.

[13] Mark de Berg, Otfried Cheong, Marc J. van Kreveld, and Mark H. Overmars. 2008. *Computational geometry: algorithms and applications, 3rd Edition.* Springer. https://www.worldcat.org/oclc/227584184

[14] Alan J. Demers, Johannes Gehrke, Biswanath Panda, Mirek Riedewald, Varun Sharma, and Walker M. White. 2007. Cayuga: A General Purpose Event Monitoring System. In *Proceedings of Biennial Conference on Innovative Data Systems Research (CIDR)*. 412–422.

[15] Yanlei Diao, Shariq Rizvi, and Michael J. Franklin. 2004. Towards an Internet-Scale XML Dissemination Service. In *Proceedings of Very Large Data Bases (VLDB)*. 612–623.

[16] Françoise Fabret, Hans-Arno Jacobsen, François Llirbat, João L. M. Pereira, Kenneth A. Ross, and Dennis Shasha. 2001. Filtering Algorithms and Implementation for Very Fast Publish/Subscribe. In *Proceedings of ACM Management of Data (SIGMOD)*. 115–126.

[17] Antonin Guttman. 1984. R-Trees: A Dynamic Index Structure for Spatial Searching. In *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, USA, June 18-21, 1984*, Beatrice Yormark (Ed.). ACM Press, 47–57. https://doi.org/10.1145/602259.602266

[18] Yuchen Li, Zhifeng Bao, Guoliang Li, and Kian-Lee Tan. 2015. Real time personalized search on social networks. In *ICDE*. 639–650.

[19] Samuel Madden, Mehul A. Shah, Joseph M. Hellerstein, and Vijayshankar Raman. 2002. Continuously adaptive continuous queries over streams. In *Proceedings of ACM Management of Data (SIGMOD)*. 49–60.

[20] Benjamin Nguyen, Serge Abiteboul, Gregory Cobena, and Mihai Preda. 2001. Monitoring XML Data on the Web. In *Proceedings of ACM Management of Data (SIGMOD)*. 437–448.

[21] Norman W. Paton and Oscar Díaz. 1999. Active Database Systems. *Comput. Surveys* 31, 1 (1999), 63–103.

[22] Miao Qiao, Junhao Gan, and Yufei Tao. 2016. Range Thresholding on Streams. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 571–582. https://doi.org/10.1145/2882903.2915965

[23] Boyu Ruan, Junhao Gan, Hao Wu, and Anthony Wirth. 2021. Dynamic Structural Clustering on Graphs. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 1491–1503. https://doi.org/10.1145/3448016.3452828

[24] Eugene Wu, Yanlei Diao, and Shariq Rizvi. 2006. High-performance complex event processing over streams. In *Proceedings of ACM Management of Data (SIGMOD)*. 407–418.

[25] Albert Yu, Pankaj K. Agarwal, and Jun Yang. 2012. Processing a large number of continuous preference top-k queries. In *Proceedings of ACM Management of Data (SIGMOD)*. 397–408.