Adaptive and Approximate Orthogonal Range Counting

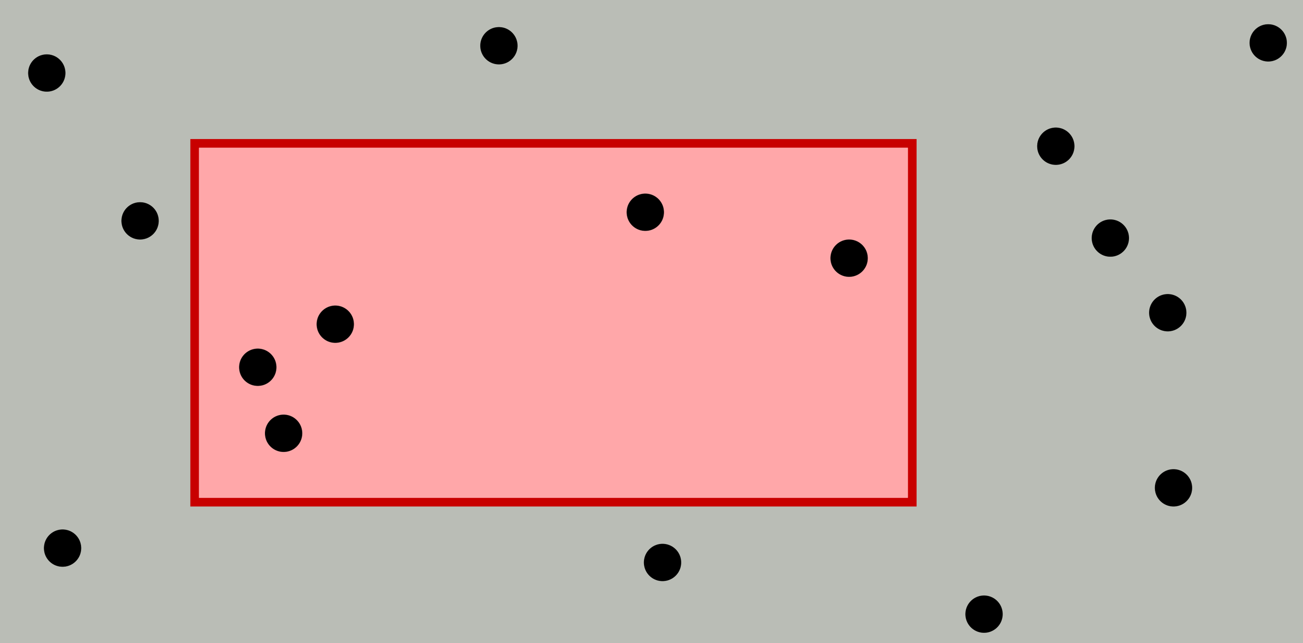
Timothy M. Chan and Bryan T. Wilkinson

David R. Cheriton School of Computer Science, University of Waterloo

UNIVERSITY OF WATERLOO

Problem

We are given a 2-D point set P of size n. Given a query rectangle Q, we must compute k, the number of points of P that lie in Q. We must create a data structure that supports a single static point set P but multiple online queries. This problem is called *static 2-D orthogonal range counting*.

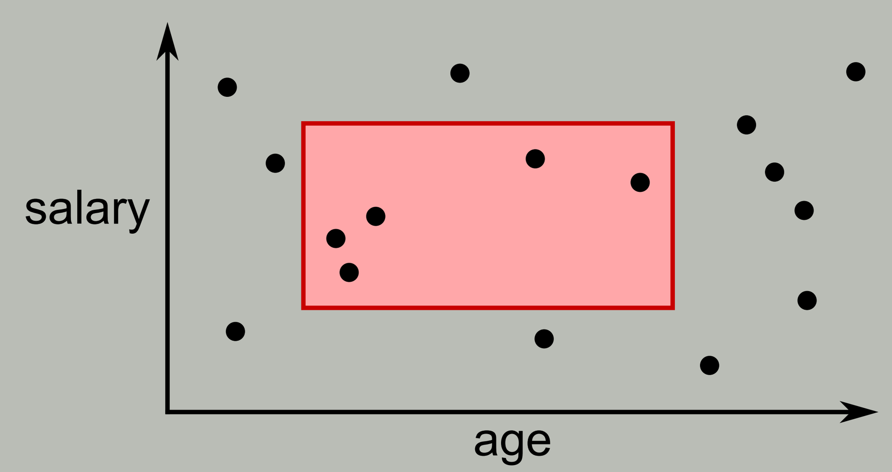


$k = 5$

Motivation

For example, an SQL query with inequality filters on two columns maps to a 2-D orthogonal range query.

```
SELECT *
  FROM employee
 WHERE age >= 20
   AND age <= 25
   AND salary >= 50000
   AND salary <= 60000
```



Model

We work under the w-bit word RAM model:

- w-bit words
- unit-cost operations on words
- fixed size universe $\{1, \ldots, U\}$

We make two standard assumptions:

- every element in the universe fits in a word (i.e., $w = \Omega(\log U)$), and
- every index into the input array fits in a word (i.e., $w = \Omega(\log n)$).

This model very closely matches modern computers operating on internal memory.

Previous Results

Reference	Space	Query Time
Bentley [Commun. ACM, 1980]	$O(n \log n)$	$O(\log^2 n)$
Willard [SIAM J. Comput., 1985]	$O(n \log n)$	$O(\log n)$
Chazelle [SIAM J. Comput., 1988]	$O(n)$	$O(\log n)$
Shi and JaJa [Tech. Report, 2003]	$O(n \log^\epsilon n)$	$O(\log_w n)$
JaJa et al. [ISAAC, 2004]	$O(n)$	$O(\log_w n)$

Pătraşcu [STOC, 2007] gives a $\Omega(\log_w n)$ lower bound on query time for data structures that use up to $n \log^{O(1)} n$ space. Thus, the data structure of JaJa et al. [ISAAC, 2004] is optimal and it seems that the problem has been solved...

A Way Forward

Chan et al. [SoCG, 2011] give a data structure for static 2-D orthogonal range emptiness (i.e., deciding whether or not $k > 0$) with efficient $O(\log \log n)$-time queries. In other words, we can count up to a maximum of 1 in $o(\log_w n)$ time. Thus, there is hope for more efficient counting data structures by parameterizing the problem on k. Note that this hope only exists under the word RAM model, as under other models, lower bounds for the emptiness problem match those for the counting problem.

New Results

We give an *adaptive* data structure that answers queries in $O(\log \log n + \log_w k)$ time. The data structure requires $O(n \log \log n)$ space. These specific bounds are important for two reasons:

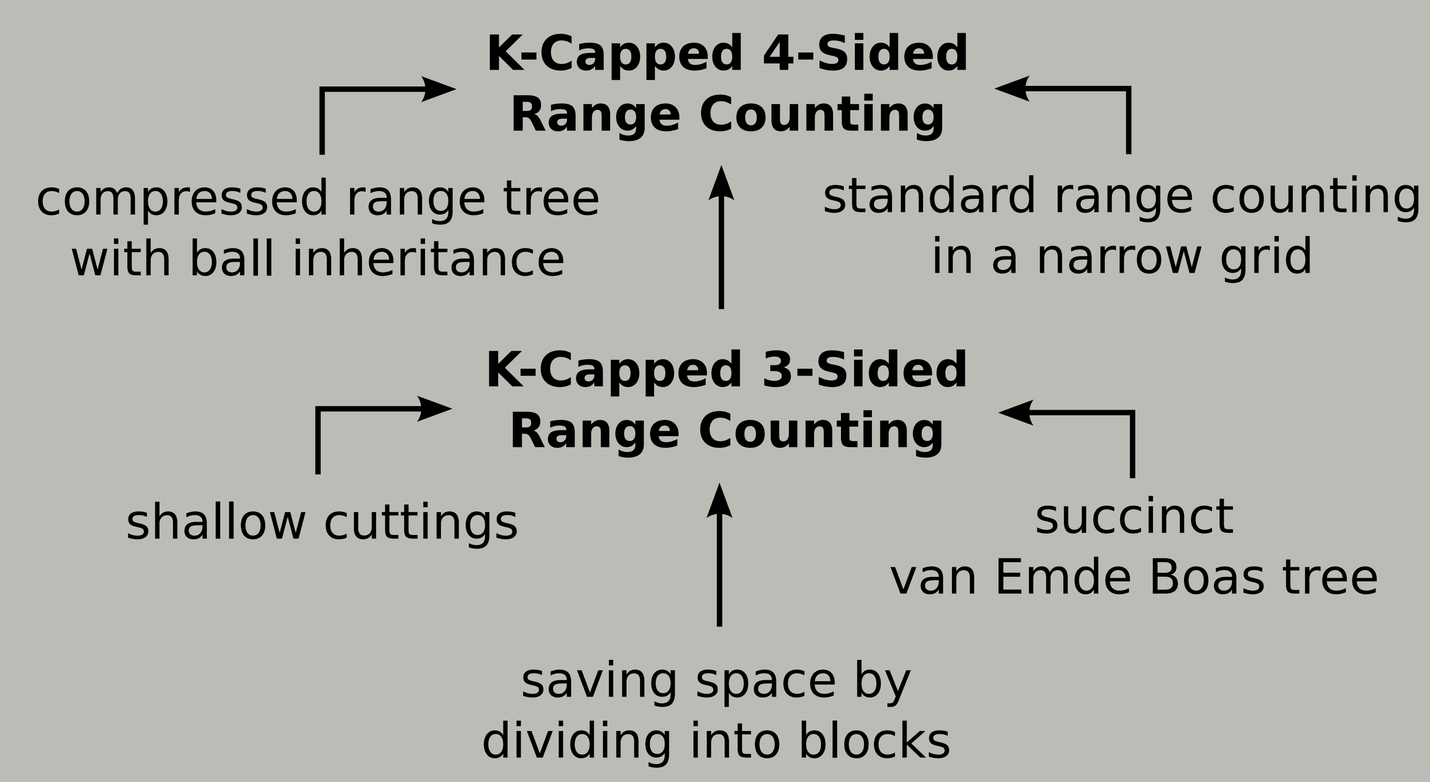
- they match the bounds of Chan et al. [SoCG, 2011] for the emptiness problem when $k = O(w^{\log \log n})$, and
- they match the lower bound of Pătraşcu [STOC, 2007] when $k = \Omega(n^\epsilon)$.

We also give data structures for *approximate* counting (when the output count can be off by a multiplicative constant factor). Our data structures match the bounds of Chan et al. [SoCG, 2011] for the emptiness problem:

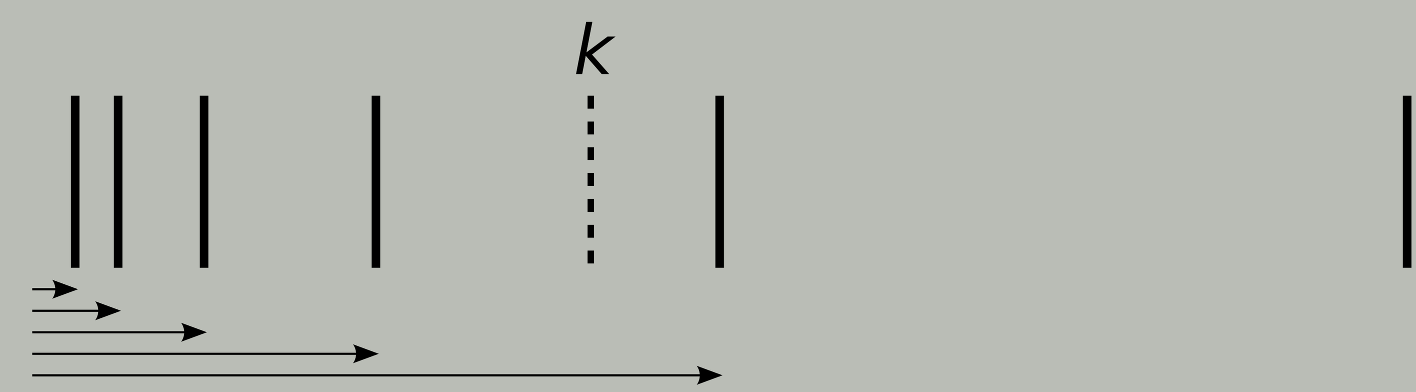
- $O(n \log \log n)$ space and $O(\log \log n)$ query time, or
- $O(n)$ space and $O(\log^\epsilon n)$ query time.

Techniques

We introduce a variant of range counting called K*-capped range counting*. In this variant, we are allowed to report failure if $k > K$. For solutions to this variant we give bounds that are parameterized on K instead of k. By combining several existing techniques in a highly non-trivial fashion, we obtain K-capped data structures with bounds that match the emptiness problem modulo extra $O(\log_w K)$ terms in their query times.



Our adaptive and approximate data structures use K-capped data structures as black boxes. Our adaptive data structure consists of $O(\log \log n)$ K-capped data structures for double-exponentially increasing values of K. An adaptive query makes K-capped queries in increasing order of K until failure is not reported. A converging geometric series keeps the sum of the $O(\log_w K)$ terms bounded by $O(\log_w k)$.



Our approximate data structure uses a K-capped data structure to handle the case where k is small and uses a standard random sampling technique to handle the case where k is large.

Open Problems

Are any of the following possible?

Problem	Space	Query Time
2-D counting	$O(n)$	$O(\log^\epsilon n + \log_w k)$
3-D counting	$O(n \log^{1+\epsilon} n)$	$O(\log \log n + (\log_w k)^2)$
2-D emptiness	$O(n)$	$O(\log \log n)$