Analyzing Different Sorting and Searching Algorithms

Zeid Ayssa
University of Texas at Austin
Dept. of Electrical and Computer Engineering
azeid@utexas.edu

Jose Martinez Garcia-Vaso
University of Texas at Austin
Dept. of Electrical and Computer Engineering
carlosgvaso@utexas.edu

Utkarsh Vardan
University of Texas at Austin
Dept. of Electrical and Computer Engineering
uvardan@utexas.edu

ABSTRACT

In this project we will analyze a variety of sorting algorithms and explore how they compare to each other in terms of performance as well as time and space complexity. Additionally, we will also show how different sorting algorithms perform with different test cases and data-set sizes. Add to that, we will explore the possibility to visualize each sorting algorithm and demonstrate how effectually different data sets from different starting point can be sorted using different algorithms.

Some of the sorting algorithms that we are considering to study are:

- Quick Sort
- Merge Sort
- Bubble Sort
- Heap Sort
- Shell Sort
- Radix Sort
- Tim Sort

1 INTRODUCTION

The sorting problem is one of the fundamental problems in Computer Science. It consists of obtaining a permutation of a sequence of numbers sorted in non-decreasing order. The problem is defined as follows[1]:

Input: A sequence of n numbers $\langle a_0, a_1, ..., a_{n-1} \rangle$.

Output: A permutation of the input sequence $\langle a'_0, a'_1, ..., a'_{n-1} \rangle$ such that $a'_0 \leq a'_1 \leq ... \leq a'_{n-1}$.

Over the times, there have been many algorithms developed to solve this problem. These algorithms employ a variety techniques, data structures and mathematical properties, which lead to some algorithms performing better than others for different input types. In our research, we focused specifically on the problem of sorting an array of integers, and we studied the performance of nine sorting algorithms when presented with multiple inputs of different sizes and makeups. We compared the performance of the following algorithms:

- Insertion sort.
- Quicksort.
- Java Collections Framework Arrays sort implementation.
- Mergesort.
- Heapsort.
- Bubble Sort.
- Shell Sort.
- Radix Sort.
- Timsort.

In the next subsections, we will introduce the previous algorithms, and we will talk about their theoretical time and space complexities. We will also go over their best, average and worst case inputs.

1.1 Insertion Sort

The insertion sort algorithm sorts the the array of integers in-place in a similar way as we would sort a had of cards. It divides the array in a left and right sides, with the goal of ending with all the numbers sorted in the left side and no numbers in the right side. We start with all the numbers in the right side, and we add them one by one to the left side in ascending order. To make sure the numbers are

inserted in the correct place, the numbers are compared from right to left with the numbers already in the left side of the array[1].

Insertion performs single pass in the collection of data. All the data on the left side of the item currently been evaluated is know to be sorted and all the data to the right is considered to be unsorted. *Figure 1* shows insertion sort in action. We start with an unsorted array (see *Figure 1 (a)*). The first item in the collection, since there is nothing to the left of 6, it is considered to be sorted. Since 4 is less than 6, we make a swap, and 4 and 6 are considered to be sorted. The rest of the numbers are unsorted. Since 5 is less than 6, we perform a swap between 5 and 6. We continue this operation for 9, since 9 is greater than 6, no swap is performed. Since 7 is less than 9, we swap 7 and 9. Finally, since 8 is less than 9, we again swap 8 and 9. In a single pass, we have sorted the entire array of data using the insertion sort (see *Figure 1 (b)*).

Figure 2 shows the time and space complexities of insertion sort.

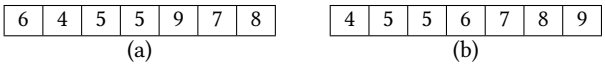

Figure 1: Insertion sort visualization. (a) shows the input array. (b) shows the array after one pass of insertion sort.

Time Complexity	
Best Case	$\Theta(n)$
Average Case	$\Theta(n^2)$
Worst Case	$\Theta(n^2)$

Space Complexity	
Worst Case	$O(1)$

Figure 2: Insertion sort time[1] and space complexities[needs citation].

1.2 Quicksort

Quick sort is a divide and conquer algorithm. It's also one of the most commonly used general purpose sorting algorithm in computer science. Since it is a divide and conquer algorithm, we will be dividing the data into smaller sets. In quick sort, the arrays are not necessarily split in half, rather a pivot value is picked based on the rules or a heuristic. Once a pivot value is picked, all the items in the array smaller than the pivot are placed at the left side of pivot, and entries to the right of the pivot are larger than the pivot. This pivot and partition operation is performed repeatedly on left and right side of partitions until all the items are sorted.

In the unsorted array of data shown in *Figure 3 (a)*, we choose 5 as the pivot. All the entries less than 5 will be moved to the left of 5, and values greater than 5 will be moved to the right of 5. Therefore, 4 and 8 are swapped resulting in *Figure 1 (b)*. Next we choose 2 as the next partition, all the values to the left of 2 are larger, and values to the right are smaller. Therefore, we need to pull 2 out of the array, and place all the values around in their appropriate location as shown in *Figure 3 (c)*. If we pick 3 as pivot, entries to the left of 3 are smaller. Therefore, it is sorted. Similarly, we repeatedly pick to pivot to the right side of 5 until all elements are sorted as displayed in *Figure 3 (d)*.

Figure 4 shows the time and space complexities of insertion sort.

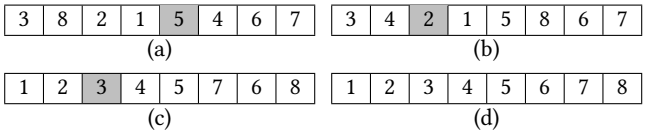

Figure 3: Quicksort visualization. (a) shows the input array. (b) shows the array after using 5 as the pivot. (c) shows the array after uisng 3 as the pivot. (d) shows the sorted array.

Time Complexity	
Best Case	$\Theta(n \lg n)$
Average Case	$\Theta(n \lg n)$
Worst Case	$\Theta(n^2)$

Space Complexity	
Worst Case	$O(\lg n)$

Figure 4: Quicksort time[1] and space complexities[needs citation].

1.3 Mergesort

Merge sort works by recursively splitting the data in half. For example, an array of 10 items would be split in the middle into two sub-arrays of 5 items each. The splitting continues until each sub-array has only one item in it. Since each sub-array has only one item in it, that sub-array is known to be sorted. At this point, the sub-arrays are merged, but the values are put together in sorted order. After each merger, the sorted sub-array doubles in size, and

this procedure continues until all the sub-arrays are merged and fully sorted.

Initially, the arrays are recursively split in half. First, the initial array (see *Figure 5 (a)*) is split into 2 sub-arrays of 4 entries each as shown in *Figure 5 (b)*. Next, the sub-arrays are split in arrays of 2 entries each as shown in *Figure 5 (c)*. Then, the resulting sub-arrays are split into 8 sub-arrays of 1 entry each as shown in *Figure 5 (d)*. Because there is only one item in the sub-arrays shown in *Figure 5 (d)*, the entries in the sub-arrays are sorted within their sub-array. For the reconstruction phase, we merge each single entry sub-array back to sub-arrays of 2 entries each by sorting them. Then, 4 and 9 are reconstructed into a sub-array as they are already sorted. 3 and 2 are reconstructed into an array with 2 and 3 in sorted order. 6 and 5 are reconstructed as 5 and 6. 7 and 8 are reconstructed in the same order. The result is shown in *Figure 5 (e)*. Similarly, we merge the resulting sub-arrays of 2 entries each back to 2 sub-arrays of 4 entries each in sorted order, which is shown in *Figure 5 (f)*. In the final reconstruction step, the 2 remaining sub-arrays are merged into a single sorted array as shown in *Figure 5 (g)*.

Figure 6 shows the time and space complexities of Mergesort.

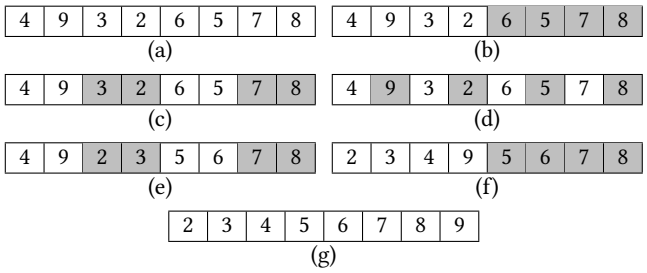

Figure 5: Mergesort visualization. (a) shows the input array. (b), (c) and (d) show how mergesort splits the initial array into sub-arrays. Different sub-arrays are distinguished by the different cell backgrounds. (e), (f) and (g) shows how mergesort joins back the sub-arrays in sorted order.

Time Complexity	
Best Case	$\Theta(n \lg n)$
Average Case	$\Theta(n \lg n)$
Worst Case	$\Theta(n \lg n)$

Space Complexity	
Worst Case	$O(n)$

Figure 6: Mergesort time[1] and space complexities[needs citation].

1.4 Heapsort

Heapsort is comparison based sorting technique based on binary heap data structure. It is similar to selection sort in terms of finding the maximum element, and placing the maximum element in the end. This process is repeated for the remaining elements in an array.

Initially, heap data structure is created using a min-heap or max-heap for the unsorted list of elements. The first element (root node) of the heap is either largest or smallest depending if it is a min or max-heap. We take out first element of the heap, and we place it at the end of the array. Then we reduce the heap size to keep the sorted element outside of the heap. We again make the heap using the remaining elements, and we pick the first element of the heap store in the end of the array. This process is repeated until we have the array completely sorted.

Consider the example shown in *Figure 7*, where we use a max-heap. After building the max-heap, we get the array shown in *Figure 7(b)*. Since 9 is the largest element (root node of the heap), it is removed from the heap, and it is swapped with the last entry of the array as shown in *(c)*. The size of the heap is reduced by one, which is shown by graying out the entries that are no longer part of the heap. Rebuilding the max-heap with the remaining entries, we obtain the array shown in *(d)*. Since 6 is the root of the heap, it is swapped with the last element of the heap, and it is removed from the heap by reducing its size by one (see *(e)*). *(f)* shows the result of rebuilding the max-heap, and *(g)* shows the result of removing the root node of the heap. Continuing this process, we reach a fully sorted array as shown in *Figure 7 (j)*.

Figure 8 shows the time and space complexities of Heapsort.

1.5 Bubble Sort

Bubble sort works by passing the data in the collection by multiple passes. The data is processed from start to end, or left to right. Starting from the first value in the collection, the value is compared to next value, if the value is larger than the next value they are swapped. The largest value is moved to the right. This comparison and swap operation is repeated for each value in the collection until no swaps are performed. Then, the array is in sorted order.

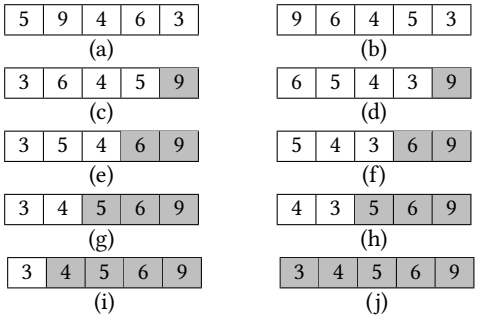

Figure 7: Heapsort visualization. (a) shows the input array. (b), shows how initial array is made into a max-heap. In (c), the root node is swapped with the last element of the array, and the heap size is reduced by one. The elements in the heap are in white, while the elements outside the heap are greyed out. The previous steps are repeated from (d) to (j), until a fully sorted array is produced as shown in (j)

Time Complexity	
Best Case	$\Theta(n \lg n)$
Average Case	$\Theta(n \lg n)$
Worst Case	$\Theta(n \lg n)$

Space Complexity	
Worst Case	$O(1)$

Figure 8: Heapsort time[1] and space complexities[needs citation].

Figure 9 shows how bubble sort works on an example array. In the 1st pass (see *Figure 9 (a)* and *(b)*), we compare 4 with 7, and swap is not performed because 4 is already smaller than 7. Next, 7 is compared and swapped with the next elements: 5 and 5 (*(b)* and *(c)*). Similarly in the third pass *(d)*, 7 is swapped with 6. As shown in *e*, the fourth pass *(e)* yields a sorted array, since no swaps are performed.

Figure 10 shows the time and space complexities of Bubble sort.

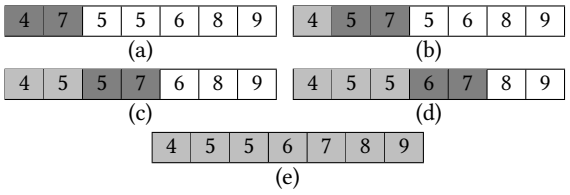

Figure 9: Bubble sort visualization. (a) shows the input array. The comparisons are shown in dark gray, and sorted elements are shown in light gray. (b) shows pass 1 result, (c) shows pass 2 result, (d) shows pass 3 result and (e) shows pass 4 result which returns the array in sorted order.

Time Complexity	
Best Case	$\Omega(n)$
Average Case	$\Theta(n^2)$
Worst Case	$\Theta(n^2)$

Space Complexity	
Worst Case	$O(1)$

Figure 10: Bubble sort time[needs citation] and space complexities[needs citation].

1.6 Shell Sort

Shell sort is a generalization of the insertion sort algorithm. Using shell sort, we compare the elements of the array that are far apart, rather than adjacent. For this purpose, we define the variable *gap*, and we compare the elements that are *gap* number of elements away from each other. This divides the initial array in a *gap* number of interleaved sub-arrays, whose elements are sorted by comparison. With every iteration, we reduce the gap between the elements being compared, and we sort the sub-arrays. When we sort in the last pass where the $gap = 1$, the array is fully sorted. In the last pass, shell sort behaves like insertion sort.

Consider the example in *Figure 11 (a)*, where we want to sort the array in ascending order. In this example, we start with a value of $gap = 2$, but the initial value of *gap* could be anything, and we reduce the value of gap by dividing it by 2 with each pass. In *(b)*, we show how the initial array is divided in interleaved sub-arrays which are shown in different colors. Next, we sort the sub-arrays individually, as shown in *(c)* and *(d)*. Now, we are done with the iteration, and we can move to the next iteration with $gap = 2/1 = 1$. This means the whole array will be considered now as shown in *(e)*. Now, we would rearrange the elements in sorted order, but the items were already sorted in our example (see *(f)*).

Figure 12 shows the time and space complexities of Shell sort.

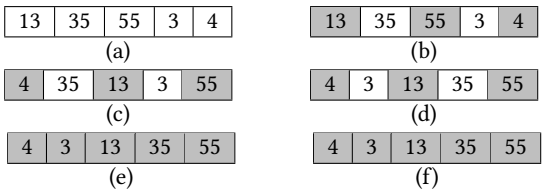

Figure 11: Shell sort visualization. (a) shows the input array. (b), (c) and (d) show how mergesort splits the initial array into sub-arrays. (e), (f) and (g) shows how mergesort joins back the sub-arrays in sorted order.

Time Complexity	
Best Case	$\Omega(n \lg n)$
Average Case	$\Theta(n (\lg n)^2)$
Worst Case	$\Theta(n (\lg n)^2)$

Space Complexity	
Worst Case	$O(1)$

Figure 12: Shell sort time[needs citation] and space complexities[needs citation].

1.7 Radix Sort

Radix sort is not a comparative integer sorting algorithm as the other algorithms that we have seen so far. It instead sorts data lexicographically. In the case that the data is made of integers, Radix sort uses the digits of the integers to sort the data, and it uses Counting sort to help with sorting.

Consider the input array shown in *Figure 13 (a)*. In order to sort elements using radix sort, we use Counting sort using the last digit of each number as the sorting keys (see *(b)*). Sorting the elements, we obtain the array shown in *(c)*. As counting sort is stable sorting algorithm, the elements with the same key will appear in the same order. In the next step, we sort the elements by counting sort using the second to last digit as key (see *(d)*), and we obtain the array shown in *(e)*. Repeating the same procedure with the first digit as key (see *(f)*), we notice that there are few numbers without any number in the first place. In this case, we pad the numbers zeroes as their key. Once we sort the first digit using counting sort, the whole array is sorted as displayed in *(g)*.

Figure 14 shows the time and space complexities of Radix sort.

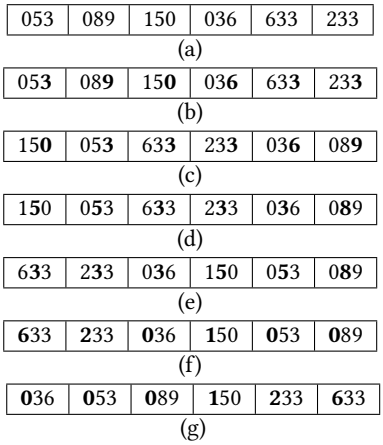

Figure 13: Radix sort visualization. (a) shows the input array. In (b), the sorting key is marked in boldface, and the resulting array after sorting is shown in (c). This procedure is repeated in (d)-(e) and (f)-(g). Notice the numbers have been padded with zeroes on the left, so that they can be sorted.

Time Complexity	
Best Case	$\Omega(nk)$
Average Case	$\Theta(nk)$
Worst Case	$\Theta(nk)$

Space Complexity	
Worst Case	$O(n + k)$

Figure 14: Radix sort time[needs citation] and space complexities[needs citation].

1.8 Timsort

Timsort is a hybrid sorting algorithm. Timsort is a combination of Insertion sort and Mergesort. We basically divide an array into sub-arrays of length RUN, where RUN is constant. Next, we apply binary insertion sort on each sub-array. Then, we merge the sub-arrays using Mergesort recursively to get the resulting sorted array. Timsort is extremely fast for nearly sorted data sequence. Timsort is also the default sorting algorithm in Java and Python, and it was implemented in 2002 by Tim Peter.

Consider the example shown in *Figure 15*, where an array of 10 elements is sorted using $RUN = 5$. *b* shows how the initial array *(a)* is divided into sub-arrays of size RUN. Using Insertion sort, we sort the sub-arrays as shown in *(c)*. Finally, the sub-arrays are merged back in sorted order using Mergesort. The resulting sorted array is shown in *(d)*.

Figure 16 shows the time and space complexities of Timsort.

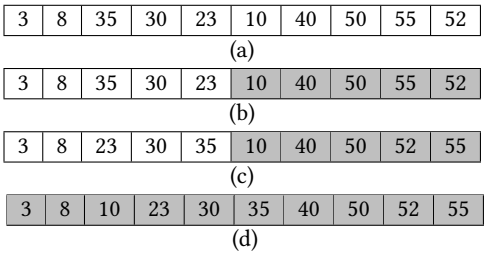

Figure 15: Timsort visualization. (a) shows the input array. (b) shows the array being split into sub-arrays of size RUN. (c) shows the sub-arrays after being sorted using Insertion sort. (d) shows the resulting sorted array after using Mergesort join back the sub-arrays in sorted order.

Time Complexity	
Best Case	$\Omega(n)$
Average Case	$\Theta(n \lg n)$
Worst Case	$\Theta(n \lg n)$

Space Complexity	
Worst Case	$O(n)$

Figure 16: Bubble sort time[1] and space complexities[needs citation].

2 IMPLEMENTATION

We have made available the source code of all implementations at the repository https://github.com/azeid/SortingAndSearchingAlgorithms.

2.1 Sorting Algorithms Correctness

As part of adding the sorting algorithms for this project, we also added correctness checks to make sure that the sorting algorithm is producing a sorted result and that the resulting array is a permutation of the original input array. This we we have confidence that the algorithm is correct. Additionally, we added smaller unit tests for each algorithm to make sure we cover the corner cases and validate that the algorithms handle unexpected inputs gracefully.

2.2 Test Cases

In order for us to compare the time complexity of different sorting algorithms, we had to come up with different test cases. We made available all the test cases used in the comparison at the repository https://github.com/azeid/SortingAndSearchingAlgorithms/tree/master/testCases.

The naming convention used was "TestCase_DataSize.txt", for example "testCases/SortedInAscendingOrderCase _1000.txt". The naming convention made it easy to identify test cases and parse the output report from the benchmark tool.

2.2.1 Test Cases Generation

Generating test cases is a tedious task if done manually; however, we automated the process of generating different test cases. We have 11 unique test cases that we ran on each algorithm. Additionally, we created the test case generation functions to generate any data size for any particular test case. Also, we tried to include test cases that cover the best, worst, and average cases for our sorting algorithms in order for the comparison to show advantages and disadvantages of different sorting algorithms based on test cases and data sizes.

2.2.2 Test Cases Generation Correctness

For every test case generator function we included unit tests to verify the correctness of the generated arrays. Given that the functions can handle any data size, we have to make sure that for large data sizes that correctness still holds. These unit tests gave us confidence that our test cases are what we expect them to be.

2.2.3 Test Cases Used

We have a total of 11 unique test cases. For each test case we created multiple data sizes starting from 100 to 10,000,000 (since 10 was excluded since it was too small for any comparison). We were not able to upload the 10,000,000 data size test cases as they were too large.

The test cases used are:

(1) Sorted In Ascending Order
(2) Sorted In Descending Order
(3) Random Order
(4) Random Order - High Numbers on First Half and Low Numbers of Second Half
(5) Random Order - Low Numbers on First Half and High Numbers of Second Half
(6) Sorted In Ascending Order - High Numbers on First Half and Low Numbers of Second Half
(7) Sorted In Descending Order - High Numbers on First Half and Low Numbers of Second Half
(8) Nearly Sorted In Ascending Order
(9) Nearly Sorted In Descending Order
(10) Same Value Array
(11) Merge Sort Worst Case (This was added since the Merge Sort worst case was not covered by our generic test cases)

2.2.4 Graphing Tool

In order to represent the data from the benchmarks, we developed a graphing utility. The tool parses the resulting data files from the benchmarks, and it produces a graphical representation of the data. The data plotted is the execution times of the studied sorting algorithms for different input sizes and input cases. The utility has

the ability to normalize the data using the results of any of the algorithms benchmarked. The data is organized by plotting the different input cases (the inputs cases mentioned in the previous section) in separate graphs. Each graph represents time of execution of the algorithms versus the size of the input data for an specific input case.

The tool was implemented using Python 3. It uses the matplotlib plotting library to draw the graphs. The source code of the utility is available in the repository of the project at the following URL: https://github.com/azeid/SortingAndSearchingAlgorithms/tree/master/graph_generator.

3 EVALUATION

3.1 Methodology

All of our implementations were written in Java. Initially we thought of measuring the performance of our implementations with a naive approach where we would measure the time either in milliseconds or nanoseconds for sorting the input array. We could for example simply use `System.currentTimeMillis()` or `System.nanoTime()`. However we quickly found out how unreliable the results would be, since we noticed really inconsistent results between runs. As Ponge[2] explains, this kind of benchmarking might be viable in programs written in statically compiled languages like C. However Java runs on a Virtual Machine and it uses *Just-in-time* compilation, so the first time the code is run it is actually being interpreted and then is compiled to native code, depending on the actual platform that is running. Furthermore, the VM tries to use all kinds of different optimization like loop unrolling, inlining functions or on-stack replacements, making it difficult to get consistent results.

We decided to use Java Microbenchmark Harness (JMH)[1] for measuring the performance of our implementations. JMH is an open source benchmarking tool part of the OpenJDK. Although it does not entirely prevent all common pitfalls and inconsistencies introduced by the JVM, it does help mitigating them.

Next step was to use the generated test cases files as inputs for our benchmarks. There are different modes to run benchmarks in JMH. We decided to measure the average time of an operation in microseconds, where an operation is sorting the array for any given algorithm and input array. In this mode, JMH considers an iteration to be a slice of time running as many operations as possible, it measures the time for each operation and averages it. In order to avoid some of the JIT inconsistencies and other JVM optimizations, JMH runs a few warm-up iterations. After that it runs, by default, 5 iterations where the results are actually recorded. For our measuring purposes we decided to run 3 five seconds warm-up iterations and 5 ten seconds actual iterations.

The overall benchmark running time was over 36 hours. This was because we had large data sizes and due to the bad performance of some sorting algorithms in their worst case such as Bubble sort. We had to exclude Insertion and Bubble sort when running the test cases with a data size of 10,000,000.

3.2 Results

TODO

We ran the benchmark suite in machine with an AMD Ryzen 1700 8-core 16-threads @ 3.6Ghz and 16Gb of DDR4 RAM @ 3,200Mhz. As a baseline we decided to run all the tests in our serial implementation of Gale-Shapley. Results are shows in Table 1. In every cell the amount of milliseconds to complete an operation of the algorithm, i.e. get a result, and the margin of error is specified and a confidence interval of 99%.

n	Best (ms/op)	Random (ms/op)	Worst (ms/op)
10	0.011 ±,0.001	0.014 ±,0.001	0.022 ±,0.001
100	0.867 ±,0.009	1.198 ±,0.003	6.595 ±,0.227
200	3.463 ±,0.064	5.423 ±,0.238	43.587 ±,0.563
1000	87.740 ±,0.628	139.563 ±,3.146	8782.987 ± 3636.943

Table 1: Serial Gale-Shapley

We then followed by running our parallel version of Gale-Shapley. Results are shown in Table 2.

Finally we ran our approach to divide and conquer in parallel. Results are shown in Table 3

[1] http://openjdk.java.net/projects/code-tools/jmh/

n	Best (ms/op)	Random (ms/op)	Worst (ms/op)
10	0.176 ±,0.011	0.177 ±,0.004	0.176 ±,0.010
100	0.379 ±,0.013	0.376 ±,0.011	1.082 ±,0.009
200	0.564 ±,0.022	0.728 ±,0.059	5.576 ±,0.106
1000	2.115 ±,0.023	3.751 ±,0.168	482.950 ±,31.277

Table 2: Parallel Gale-Shapley

n	Best (ms/op)	Random (ms/op)	Worst (ms/op)
10	0.168 ±,0.003	0.169 ±,0.007	0.174 ±,0.013
100	0.691 ±,0.007	1.008 ±,0.035	1.999 ±,0.025
200	2.153 ±,0.034	3.773 ±,0.150	8.559 ±,0.304
1000	43.971 ±,0.752	74.536 ±,2.207	401.745 ±,7.819

Table 3: Parallel Tseng-Lee

REFERENCES

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms.* The MIT Press, Cambridge, MA, USA, 3rd edition, 2009.

[2] Julien Ponge. Avoiding benchmarking pitfalls on the jvm. https://www.oracle.com/technetwork/articles/java/architect-benchmarking-2266277.html, July 2014. [Online; posted July-2014].