Finding clusters

Astro 585 Presentation

Moupiya Maji

4/30/2014

Outline

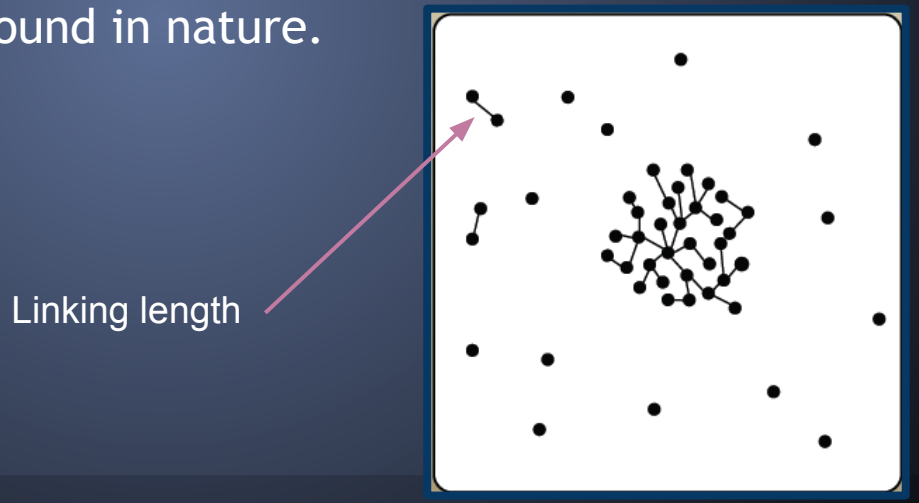
- The problem
- A solution - *Friends-of-friends*
- Basic Implementation
- The issues
- Union find implementation
- The issues
- Lessons learned

The problem

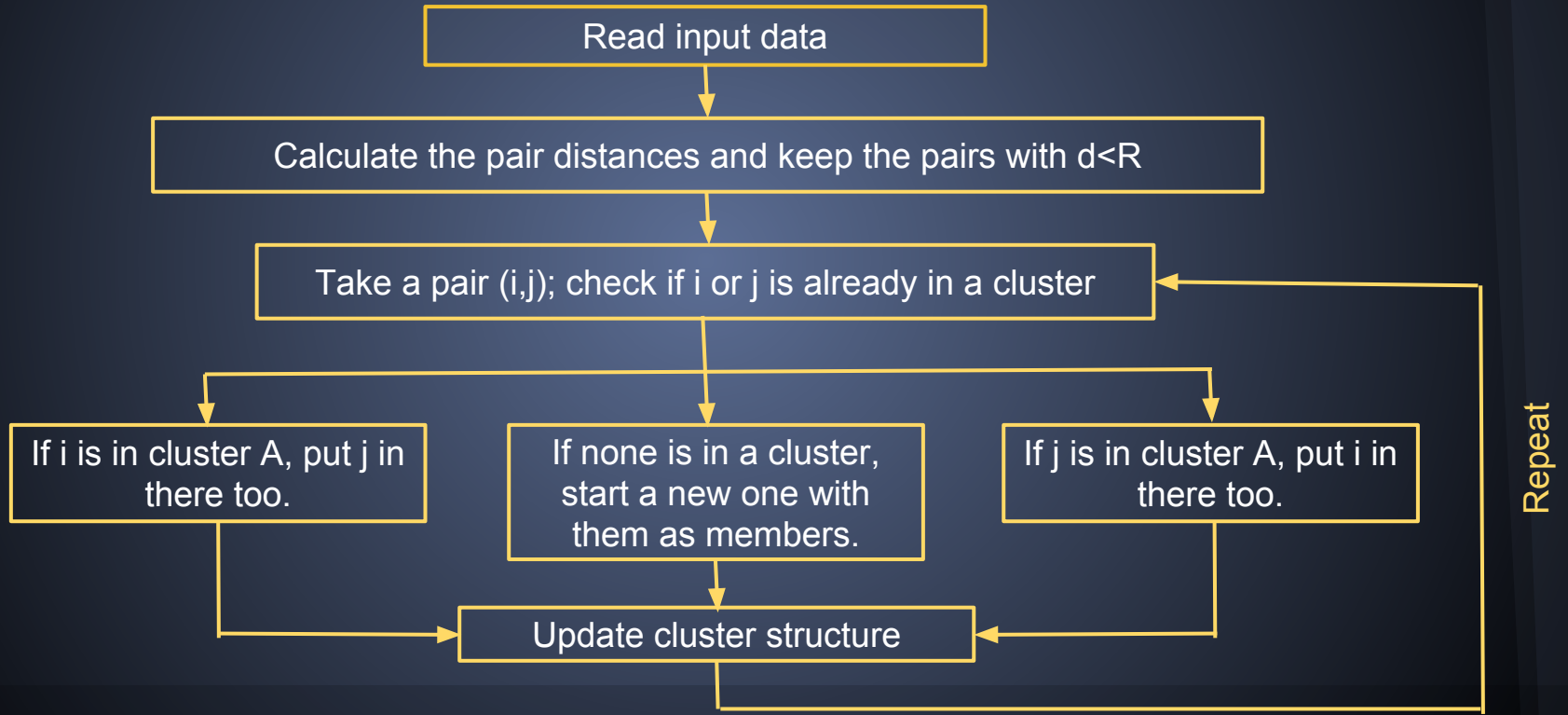
- Given a set of data points, can we find out the clusters?
- A classic problem in data mining. Data could be spatial positions or some properties of the objects.
- There are many algorithms for cluster finding which differ in their notion of clusters and in their methods of finding them. The defining factor could be small distance, overdensity or statistical distributions.

Friends-of-friends

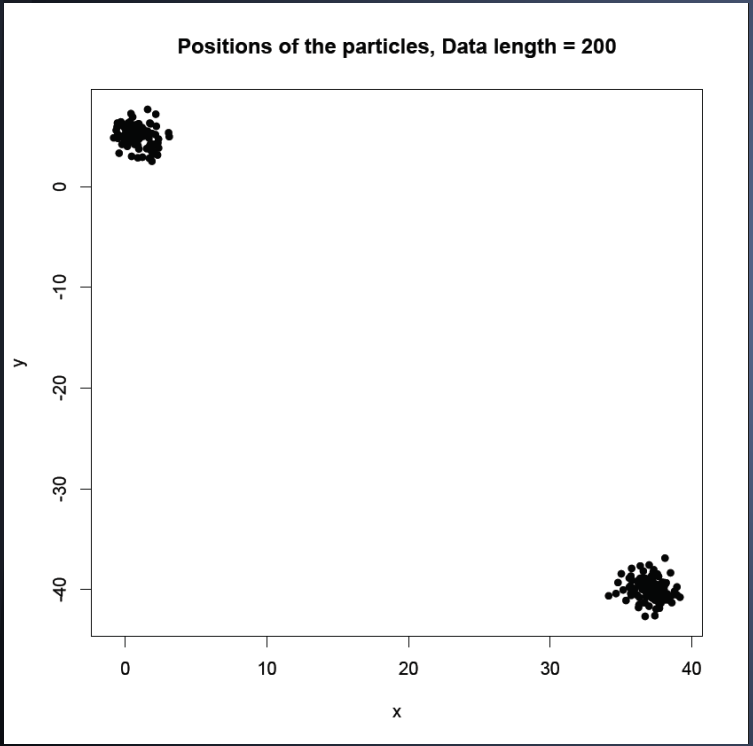
- If two particles are within a predefined distance, called *Linking Length*, they belong in the same cluster. Carry that for all the particle pairs to find clusters.
- The choice of linking length can be arbitrary. Generally, it could be typical size of the cluster found in nature.

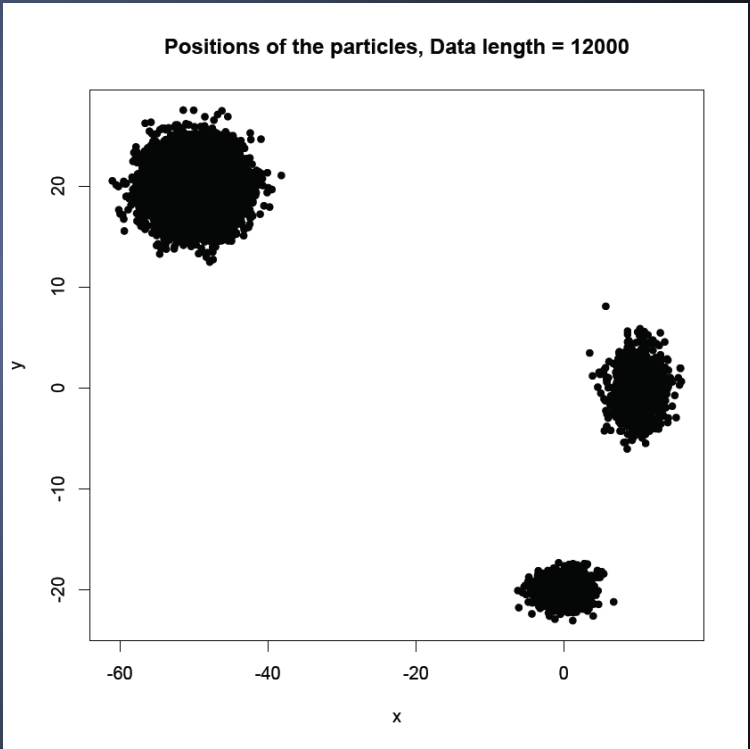


Basic implementation: algorithm

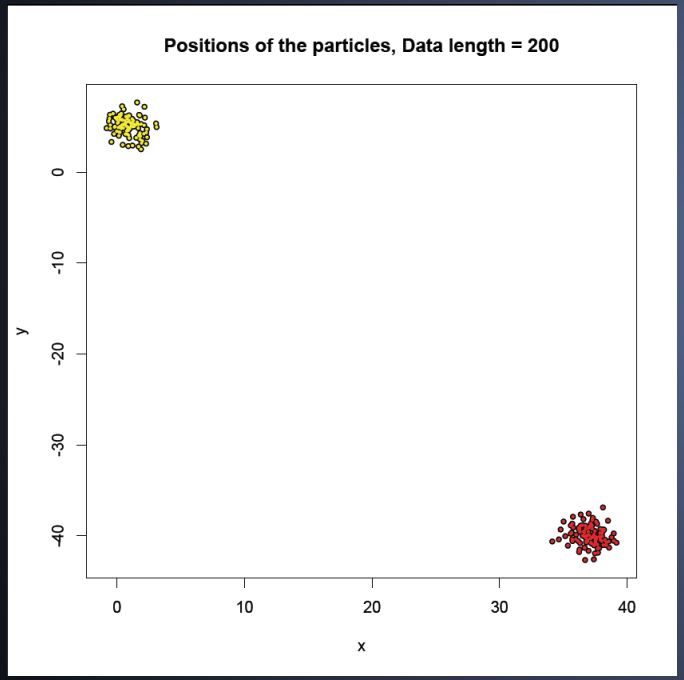


Input (positions)

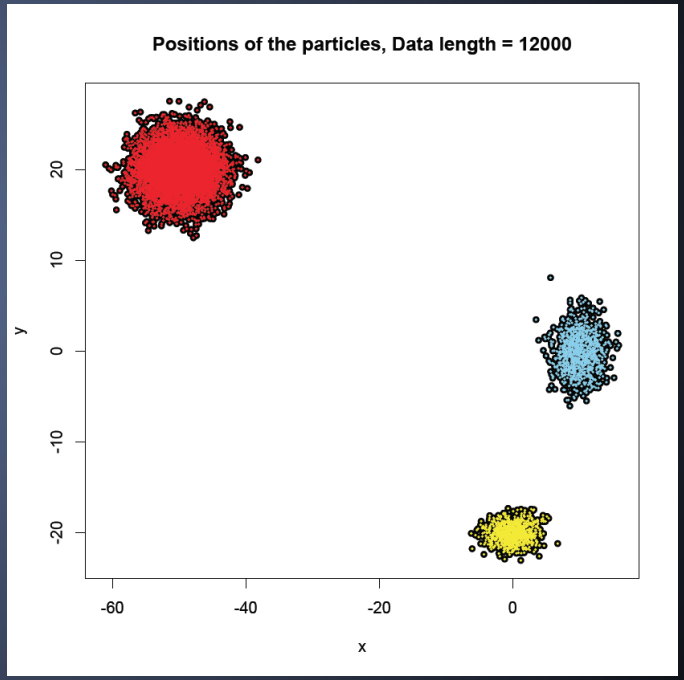




Output (clusters)

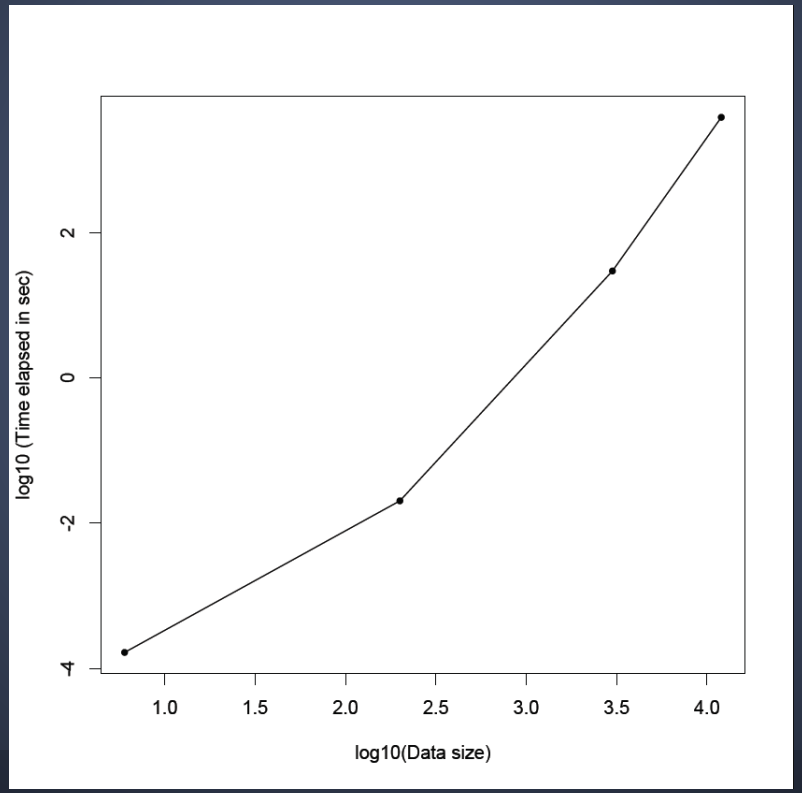


No. clusters found =2

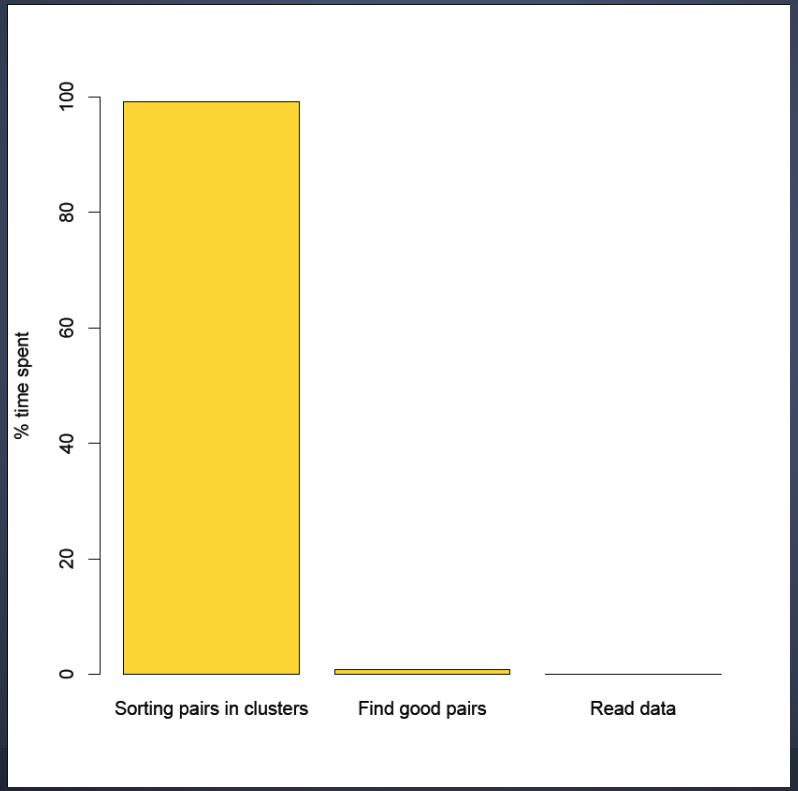


No. clusters found =3

Performance in serial version



Profiling

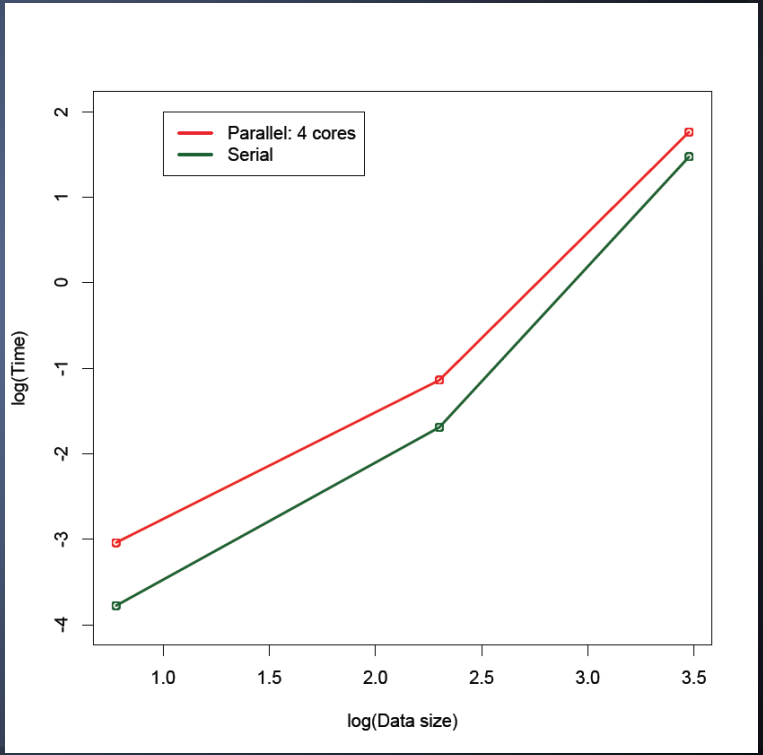

% time spent
0
20
40
60
80
100
Sorting pairs in clusters
Find good pairs
Read data

Problems in parallelizing

Every step depends on all previous steps.

Parallelizing the ‘find’ section.

Requires synchronization → Slow.



The union find method

Finding and connecting the particles much easier.

The union find method

Finding and connecting the particles much easier.

Eligible pairs

0 1
0 2
1 3
4 5
5 6
4 7

Array of all possible clusters

Indices	Values
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7

The union find method

Finding and connecting the particles much easier.

Eligible pairs	Array		Union
0 1	0	0	
0 2	1	1	0
1 3 →	2	2	0
4 5	3	3	0
5 6	4	= 4	
4 7	5	5	4
	6	6	4
	7	7	4
	Indices	Values	

The union find method

Finding and connecting the particles much easier.

Eligible pairs

0 1
0 2
1 3
4 5
5 6
4 7

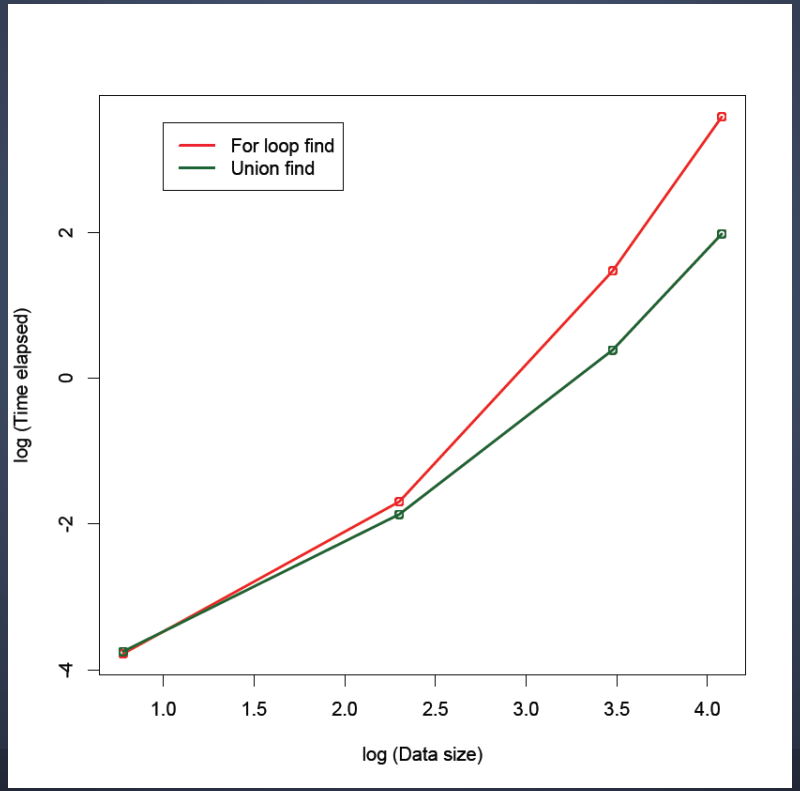
Union Array

Indices	Values
0	0
1	0
2	0
3	0
4	= 4
5	4
6	4
7	4

Find Clusters

2

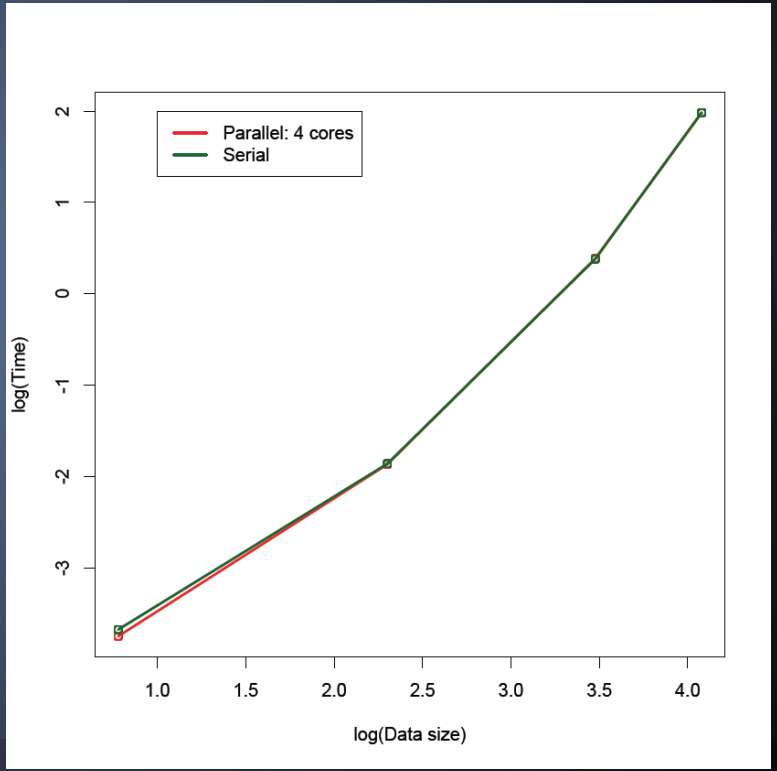
Performance in serial



Performance in parallel

Here too, steps depend on previous steps.

Parallelized the calculate pair distance part instead. No significant improvement.



Summary and tips

- Implementing a good algorithm is important.

- Parallelizing code with dependencies is difficult.
- Future plan : parallelize the union code.

Openmp tips:

- mac can limit the number of times omp for can be called inside a for loop. Could be handled with increasing stacksize.
- Array sizes inside openmp could be limited by the stacksize. Make global or dynamic.
- Vectors/ dynamic arrays are not handled well in parallel.