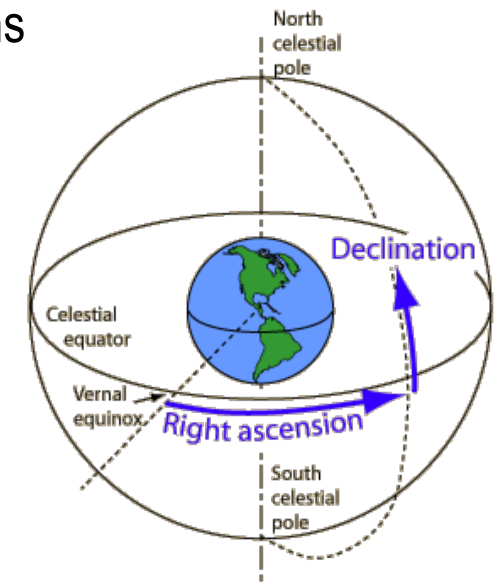
Assignment: Computing the two-point angular correlation function

- The two-point angular correlation function (TPACF) is a statistical measure of the spatial distribution of galaxies in the universe
  - a statistical measure of to what extent galaxies are randomly distributed in the universe, or lumped together
- Used to test the Lambda Cold Dark Matter model (ΛCDM), which proposes that the energy density of the universe is dominated by two unknown components
  - dark energy, which drives the expansion of space
  - dark matter, which drives the formation of structures in universe
  - together, these two account for about 96% of the total matter-energy in universe (the remaining 4% is "normal" baryonic matter, i.e. protons and neutrons)
- Dark energy or matter can not be directly observed, only its effects on other bodies
  - for instance gravitational force

Input data

- The code analyses the statistical properties of the distribution of galaxies in the universe
- The input consists of a number of position of galaxies on a celestial sphere
  - the redshift of the observations are between 0.345 and 0.355, so they are all (roughly) at the same distance from the earth
  - the observations are located on a sphere centred in earth
  - therefore, the angular separation between two observations gives the distance between the galaxies
- The input files consist of a list of coordinates in the equatorial coordinate system
  - the equatorial coordinates are pairs of right ascension and declination (ra, dec)
  - these are converted to Cartesian coordinates (x, y, z) on a unit sphere

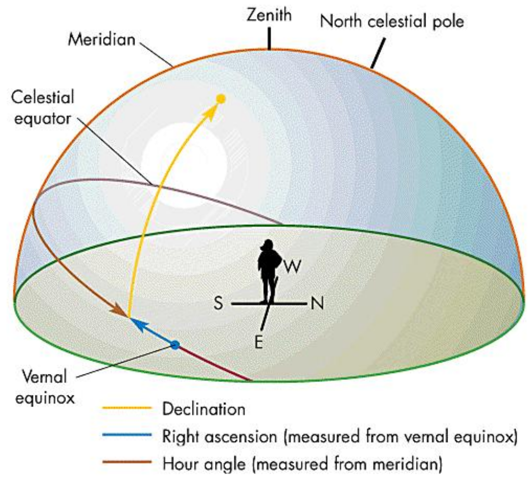


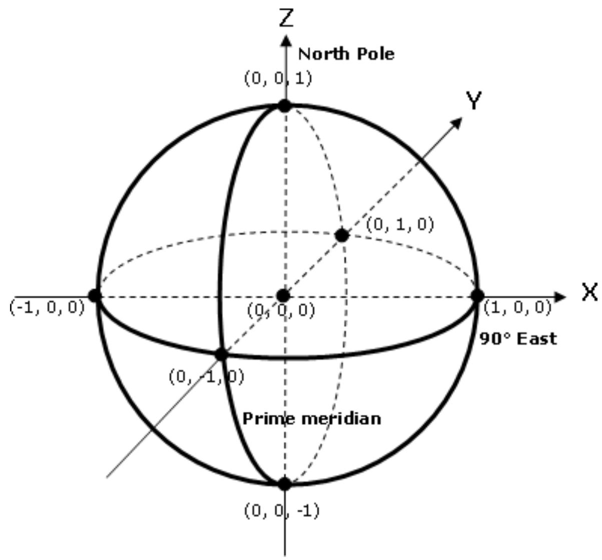
Converting to rectangular coordinates

- The input values are pairs of equatorial coordinates *(ra, dec)*
- These are converted to a rectangular coordinate system *(x, y, z)*

$$phi = ra * \pi/180.0$$
$$theta = (90.0 - dec) * \pi/180.0$$
$$x = \sin(theta)*\cos(phi)$$
$$y = \sin(theta)*\sin(phi)$$
$$z = \cos(theta)$$

- Computing the angle θ between two observation will be done with the rectangular coordinates



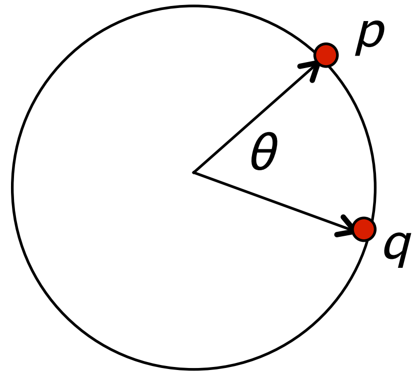


Computing angles

- Because the observations are all at the same distance from earth, the angle between two galaxies is a measure of the distance between the observations
  - two observations with a small angular separation are close to each other
- The angle between two observations $p=(x_p, y_p, z_p)$ and $q=(x_q, y_q, z_q)$ is

$$\theta = arccos(p \bullet q) = arccos(x_p * x_q + y_p * y_q + z_p * z_q)$$

where (x_p, y_p, z_p) are the rectangular coordinates of an observation p on the unit sphere



- The program computes the angle between all pairs of galaxies in two sets of input data
  - the distribution of the angles are recorded in a histogram
  - the histogram shows how many pairs of galaxies there are with a certain range of angular separation

Histogram calculation

- The program calculates the histogram distribution of angular separations between 0 and 64 degrees with a resolution of 0.25 degrees
  - the number of bins in the histograms is 4∗64 =256

- Histogram
  - calculate how many observation pairs have an angle in the intervals [0, 0.25) , [0.25, 0.5) , [0.5 – 0.75) , . . . , [63.75, 64)

5	9	3	6	15	.	.						
0	0.25	0.5	0.75	1	1.25	1.5	. . .					63.75

Åbo Akademi University

Data sets

- We have two data sets
  - the real data D consists of real observations of galaxy coordinates
  - the random data R consists of randomly generated galaxy coordinates

- The idea is to compare the real observations to a randomly generated set and calculate a statistical measure (TPACF) which tells us if the real galaxies are more lumped together in space than in a random distribution
  - if they are, this can be seen as an evidence of the gravitational forces caused by cold dark matter, which cause more attractive forces than what can be explained by the known visible mass of the galaxies

Constructing histograms

- Three histograms are calculated
  - DD(θ) : the number of observation pairs (p,q) with angular separation θ where both observations p and q are from the set of real observations
  - DR(θ) : the number of pairs (p,q) with angular separation θ where observations p are from the set of real observation and q are from the randomly generated set
  - RR(θ) : the number of pairs (p,q) with angular separation θ where both observations p and q are from the set of randomly generated coordinates

- The histogram computations have a complexity of $O(N^2)$ where N is the number of observations
  - we compute the angle for all coordinate pairs

Optimizing histogram calculations

- When the histograms *DD* and *RR* are calculated, we notice that the angle θ between two observations *i* and *j* is symmetric
  - the angle between observations number *i* and *j* in the input data is the same as the angle between observations *j* and *i*
  - both observations are taken from the same data set
  - we only need to do this calculation once
  - need only to calculate the angle for $n(n-1) / 2$ pairs of observations
- For the histogram DR this is not the case
  - the observations are taken from different data sets, so the pair (i, j) is not the same as the pair (j, i)
  - for the DR histogram we have to calculate the angle for all $n(n-1)$ pairs
- For the *DD* and *RR* histograms all angles between an observation and it self (i.e. the angle between observation pairs (i,i)) is always 0
  - no need to compute this, can instead directly add the total nr. of observations to bin number 0 in the histogram

Calculating the correlation

- The two-point angular correlation (TPACF) ω for angle θ is defined as

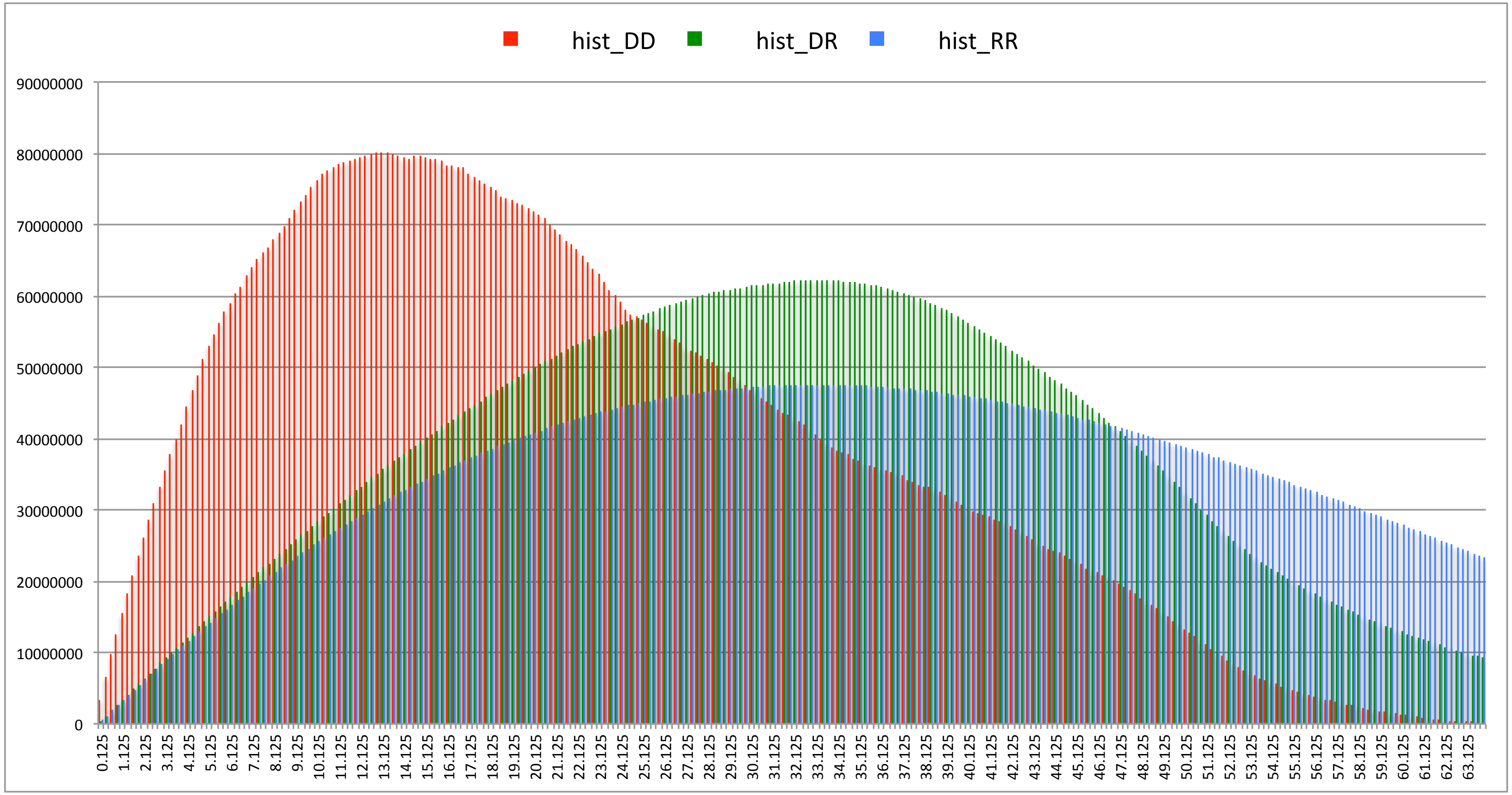
$$\omega(\theta) = 1 + \left(\frac{N_{random}}{N_{real}}\right)^2 \frac{DD(\theta)}{RR(\theta)} - 2\left(\frac{N_{random}}{N_{real}}\right)\frac{DR(\theta)}{RR(\theta)}$$

  where
  - N_{real} and N_{random} are the number of galaxies in the real and random data sets, respectively
  - $DD(\theta)$, $DR(\theta)$ and $RR(\theta)$ are the histogram values for the angle θ

- A positive value of $\omega(\theta)$ indicates that there are more galaxies with an angular separation θ than expected in a random distribution
- A negative value of $\omega(\theta)$ indicates that there are less galaxies with an angular separation θ than expected in a random distribution
- If $\omega(\theta) = 0$ the distribution of galaxies is random

Results

- The result of the computation is a table with the ω–values and histogram counts for angle intervals from 0 to 64 degrees, with an interval length of 0.25



Assignment

- Implement a parallel program using MPI to compute the two-point angular correlation as described here
  - the program should produce the same results as the sequential implementation
- As a starting point, a sequential implementation (written in C) is given
  - the program is available in a directory */export/courses/PP2014/galaxyz* on Asterope
  - written in C, about 250 lines of code
- Input files with data sets (observed and random) are also available
  - *file*.txt contains real observations
  - *file*_rand.txt contains randomly generated data

File	Nr of observations	Seq. runtime
small	20000	30 sec.
medium	100000	620 sec. (~10 min.)
large	430932	120000 sec. (~30 h.)

Testing your solution

- The program takes three file names as arguments: two input files (one with real and one with random data) and one output file
  - **Ex:** *srun –n 1 galaxyz small.txt small_rand.txt small_result.txt*
- The result is written to the output file
  - for each bin: the value of ω and the counts for the *DD*, *DR* and *RR* histograms
  - the total number of observations in each histogram is also printed out
- Check carefully that your implementation produces the same results as the sequential program
  - in particular, check the values of the DD, DR and RR histogram count
- The parallel implementations should scale up to at least 48 cores
  - test your programs with different data sizes and numbers of processors
- Document your solution
  - describe how you decompose the problem, how you implement it, the results and the scalability of your solution