Problem 1

Vectors in two coordinate systems were generated. As can be seen in Fig. 1, vectors with Cartesian coordinates (x, y) distributed uniformly in $[-1, 1]$ form a highly non uniform distribution of angles to the X-axis. On the other hand, if the Polar coordinate system (r, θ) is used and θ is chose uniformly at random from $[-\pi, \pi]$, the resulting vectors form a uniform distribution of angles to the X-axis.

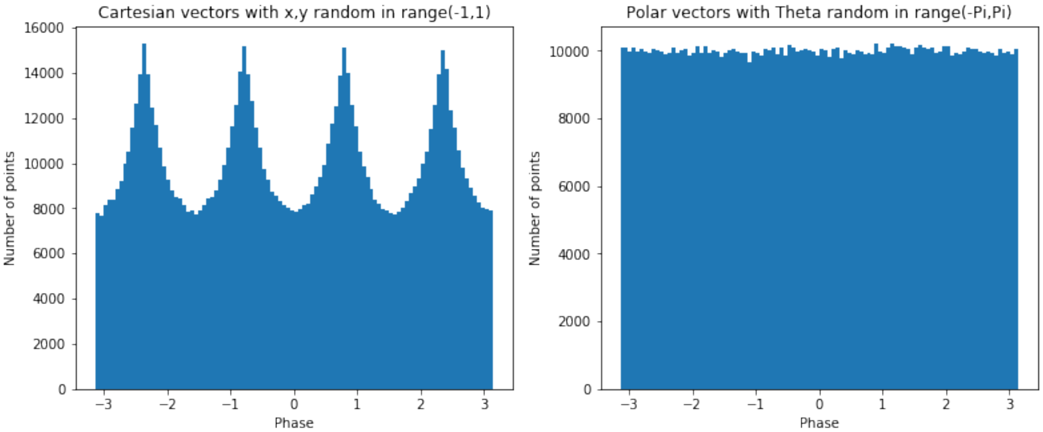

Figure 1: Distribution of angle to the X-axis in "random" vectors in Polar and Cartesian coordinate systems comparison.

Problem 2

From each machine $M_i, i \leq t$, take a small random sample $\hat{d}_i$ of the data available to the machine, call it d_i, and send it to some fixed node M_0. M_0 receives samples from each machine and based on their union is able to obtain an estimate, call it $\hat{D}$, of the distribution of the whole dataset D. M_0 then computes bin sizes such that approximately equal amounts of data from $\hat{D}$ fall into each bucket (total t buckets) and distributes this information across the t machines. Each machine M_i then should apply this bucketing as computed by M_0 to the full d_i available to it and create key-value pairs with keys equal to bucket id's and values being the data elements.

Problem 3

Part 1

The following implementation of JS computation takes (depending on hardware) in the region of 70 seconds to compute Jaccard Similarity for 5000 pairs of sets. It would take around 800 days to compute Jaccard Similarity for $\binom{10^6}{2}$ pairs of sets this way.

Algorithm 1: Computing Jaccard Similarity Directly.

```
def js(set1,set2):
    I = sum([1 for item in set1 if item in set2])
    U = len(s1) + len(s2) - I
    return I/U
```

Part 1

The distribution of JS between pairs of sets as approximated by a random sample from the dataset can be seen in Fig. 2. Overlayed in red is the Locality Sensitive Hashing S-curve for banding parameters r=5, b=7 is displayed. The maximum slope of the curve is observed at the position indicated by the green dot. The blue dot on the red curve corresponds to the value of the S function at JS=0.85.

$$b \sim \frac{klog(1/\tau)}{log(klog(1/\tau))}, \quad r = \frac{k}{b} \tag{1}$$

These values of r and b were tested on the sample of the data and it was calculated that generating minhash signatures for the full data set should around 30 minutes and hashing the bands - 15 minutes. This process generated on the order of 1000 colliding pairs of sets once the values of r and b were tuned, which took another 3 minutes to check directly using Algorithm 1.

Computing many more than ∼30 hashes per set is computationally expensive and frequently failed hard on the machine utilized for this experiment, so the value of k was chosen to be below 30. For threshold value $\tau \sim 0.85$, and $k \sim 25$ Eq. 1 produces $r = 3, \ b = 7$ (see Fig. 2). However, from empirical observation, these values lead to high rates of false positives. By observing the behaviour of the S-curve as r and b change, $r = 12$ and $b = 2$ were produced and as seen in Fig. 3 they result a much stricter filter. Threshold $\tau = 0.94$ for these values was calculated using Eq. 2.

$$\tau = \left(\frac{1}{b}\right)^{b/k} \tag{2}$$

This resulted in obtaining two pairs of sets with Jaccard Similarity ≥ 0.85. r and b were modified to $r = 12$ and $b = 3$ ($\tau = 0.91$) to increase the number of collisions (see Fig. 4).

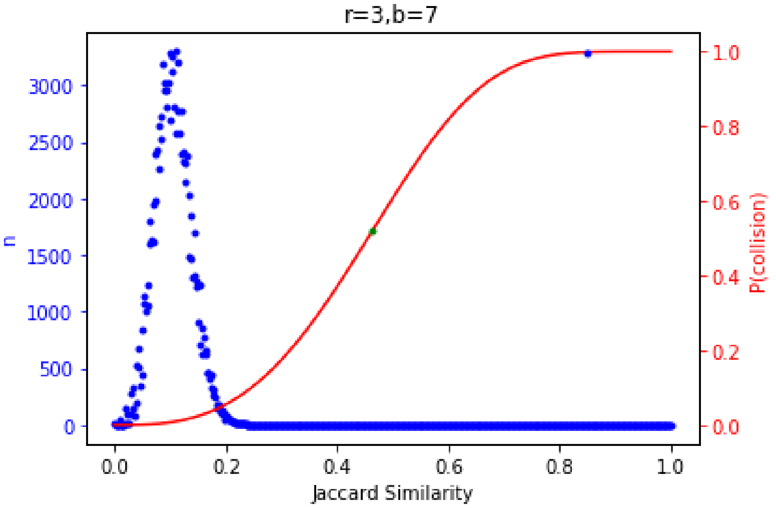

Figure 2: Distribution (blue) of true Jaccard Similarities between all pairs of sets in a random sample of the dataset and the LSH S-curve (red) given by 7 bands of 3 minhashes depicting likelyhood of collision with respect to true Jaccard Similarity. The maximum slope of the curve is observed at the position indicated by the green dot. The blue dot on the red curve corresponds to the value of the S function at JS=0.85.

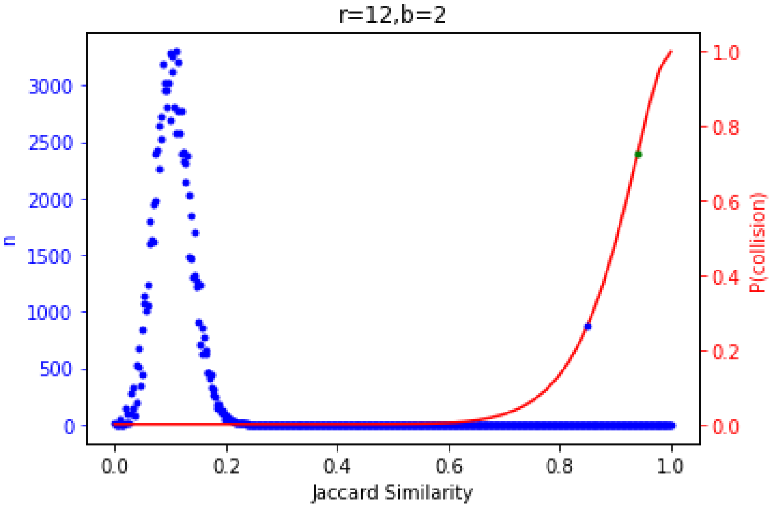

Figure 3: Distribution (blue) of true Jaccard Similarities between all pairs of sets in a random sample of the dataset and the LSH S-curve (red) given by 2 bands of 12 minhashes depicting likelyhood of collision with respect to true Jaccard Similarity. The maximum slope of the curve is observed at the position indicated by the green dot. The blue dot on the red curve corresponds to the value of the S function at JS=0.85.

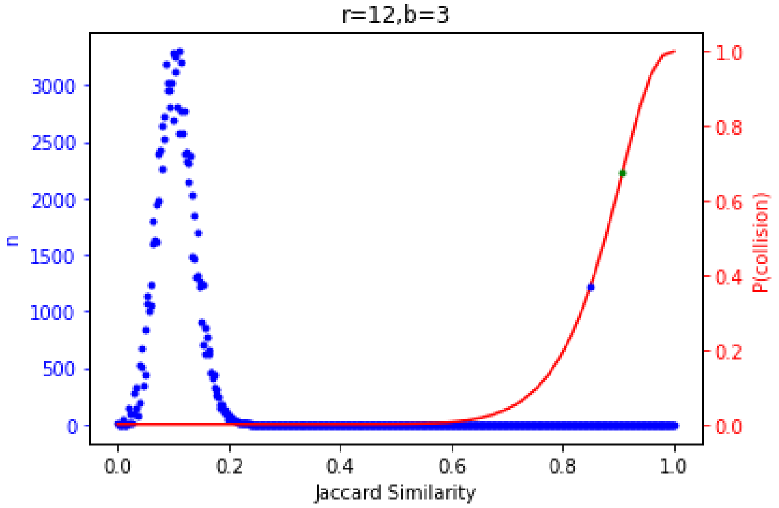

Figure 4: Distribution (blue) of true Jaccard Similarities between all pairs of sets in a random sample of the dataset and the LSH S-curve (red) given by 3 bands of 12 minhashes depicting likelyhood of collision with respect to true Jaccard Similarity. The maximum slope of the curve is observed at the position indicated by the green dot. The blue dot on the red curve corresponds to the value of the S function at JS=0.85.

The line numbers of the sets with Jaccard Similarity ≥ 0.85 can be seen in the table below with the corresponding values of JS.

Line Numbers	Jaccard Similarity
24946, 90149	0.892
104963, 42121	0.905
124849, 77683	0.864
24946, 90149	0.892
104963, 42121	0.905