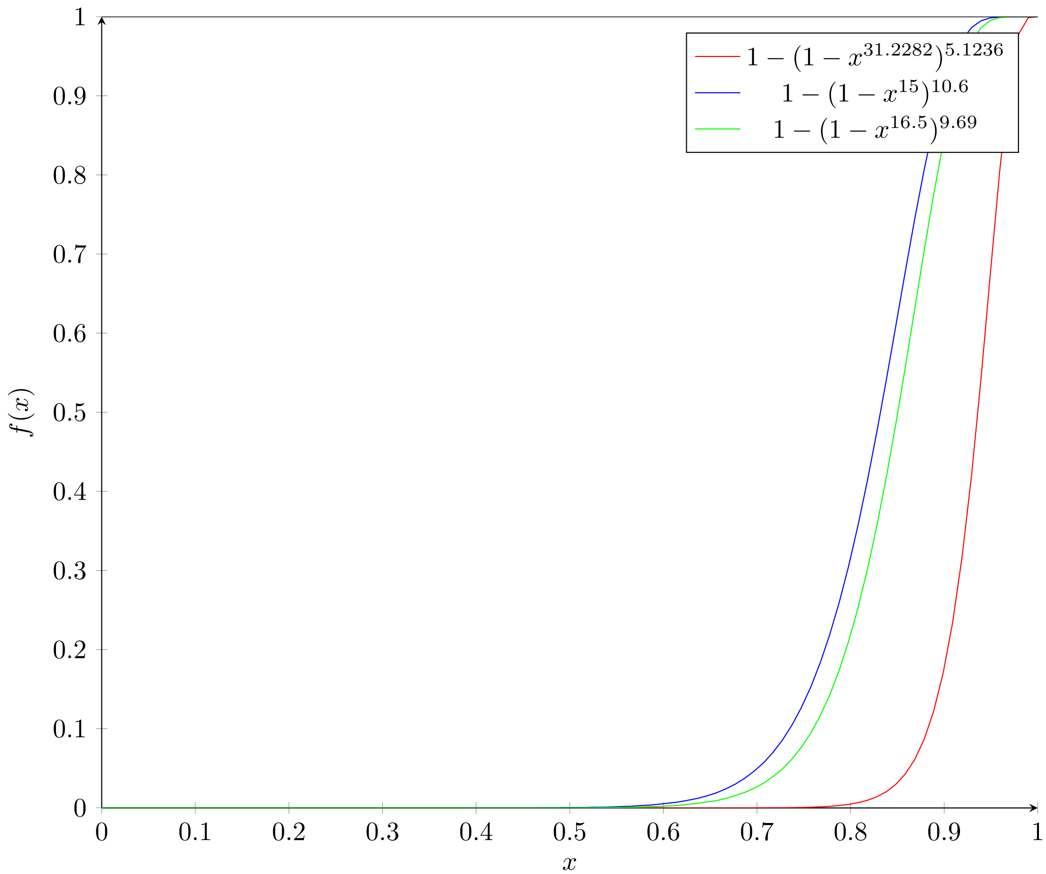
Choosing r,b

Consider computing an LSH using t = 160 hash functions. We want to find all object pairs which have Jaccard similarity above $\tau = .85$

A: (15 points) Use the trick mentioned in class and the notes to estimate the best values of hash functions b within each of r bands to provide the S-curve $f(s) = 1 - (1 - s^b)^r$, with good separation at τ . Report these values.



The value for r = 9.69 and b = 16.5 create the above curve and pass through the point (.85,.5). r * b = t $\Rightarrow$ 9.69 * 16.5 $\approx$ 160. As you can see from the equations above, r = 9.69 and b = 16.5 make up the best values.

B: (15 points) Consider the 4 objects A, B, C, D, with the following pair-wise similarities. Using your choice of r and b and f($\cdot$), what is the probability of each pair of the four objects for being estimated to having similarity greater that $\tau = 0.85$? Report 6 numbers. (Show your work.)

$1-(1-.10^b)^r = f(s)$
$1-(1-.75^{16.5})^{9.69} = .081$
$1-(1-.25^{16.5})^{9.69} = 1.12806 * 10^{-9}$
$1-(1-.35^{16.5})^{9.69} = 2.90701 * 10^{-7}$
$1-(1-.10^{16.5})^{9.69} = 3.06425 * 10^{-16}$
$1-(1-.40^{16.5})^{9.69} = 2.63217 * 10^{-6}$
$1-(1-.87^{16.5})^{9.69} = .641592$

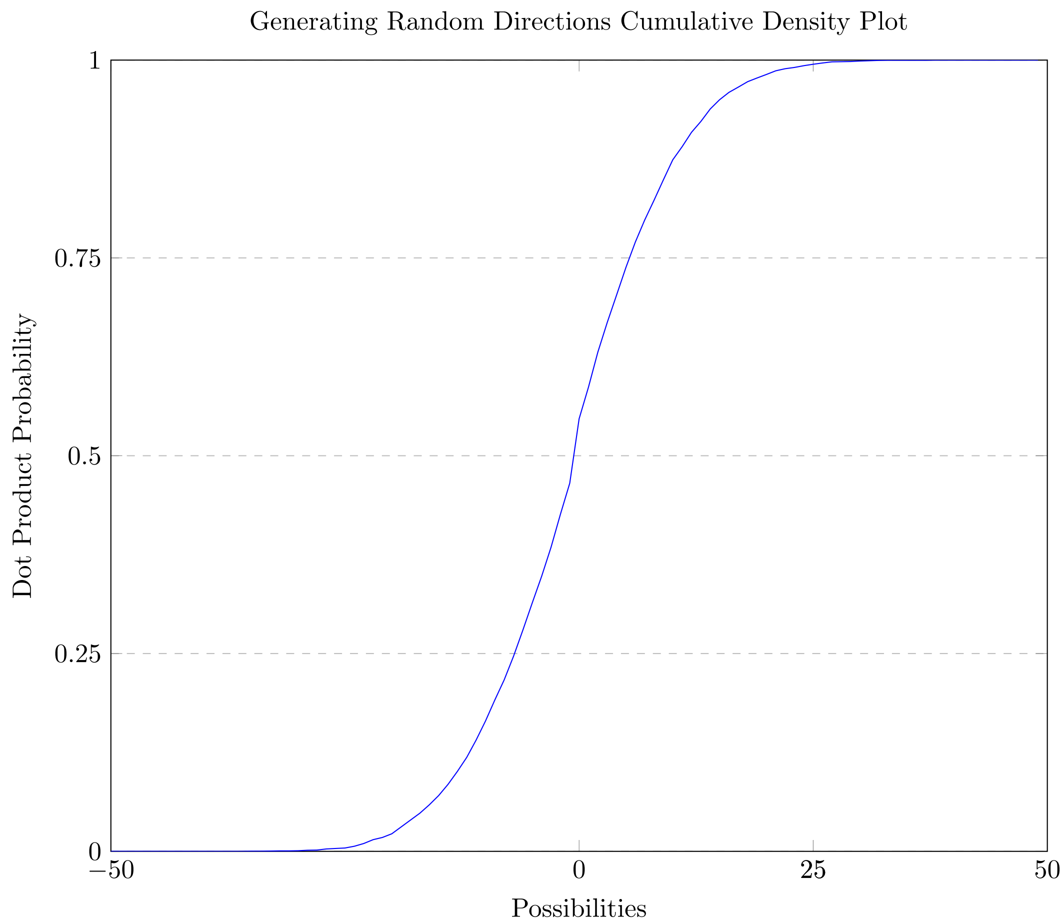
Generating Random Directions

A: (10 points) Describe how to generate a single random unit vector in d = 10 dimensions using only the operation u $\leftarrow$ unif (0, 1) which generates a uniform random variable between 0 and 1. (This can be called multiple times.)

<0.13506642406271607, 1.4369942896931038, 0.1903511525261229, -1.7537602542415558, 0.16862812390764476, 1.182598203233134, 1.2346676734218769, -0.10414957981418553, 0.357726500505956, -0.10704248845019695>

To generate a single random unit vector in d = 10 dimensions you get a program that can generate random numbers between 0 and 1. You then generate a u1 and u2 with your random number generator and you use the given equations to generate 2 numbers of the vector at the time.
Equations: y1 = $\sqrt{-2 * ln(u1)} * cos(2 * \pi * u2)$
y2 = $\sqrt{-2 * ln(u1)} * sin(2 * \pi * u2)$

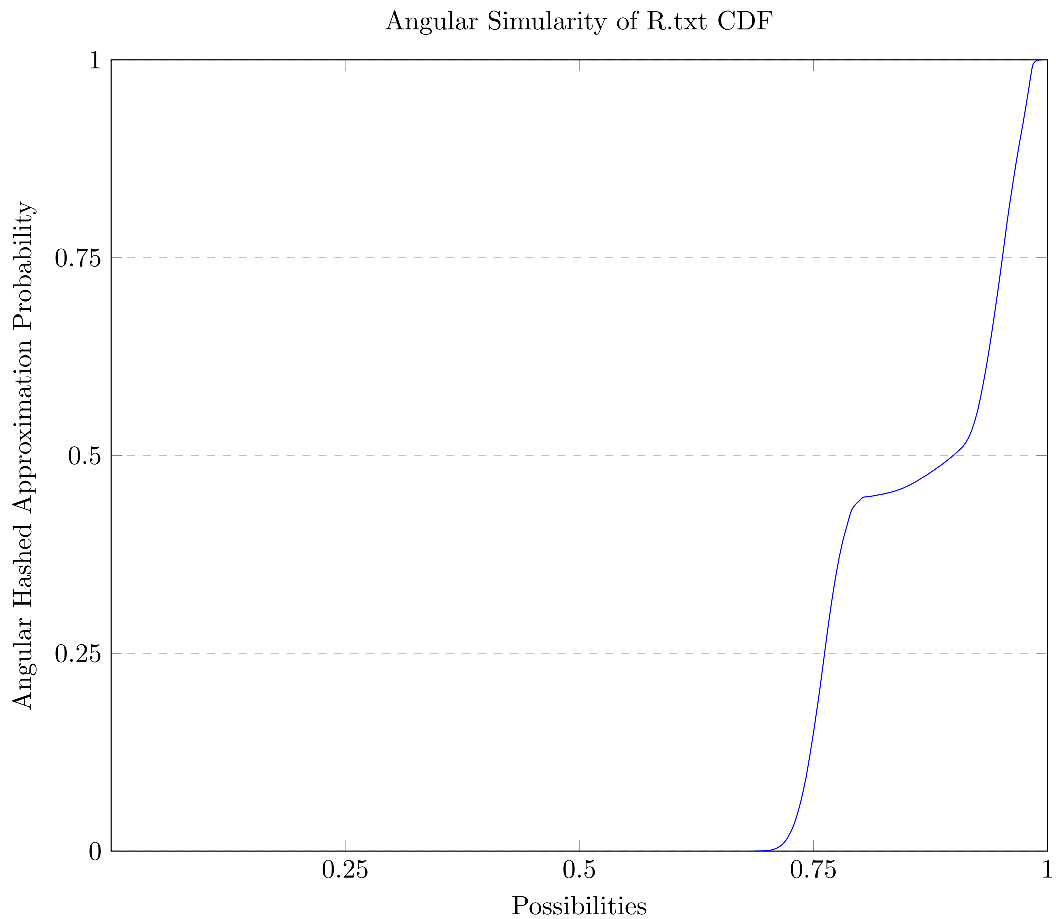
B: (20 points) Generate t = 160 unit vectors in R^d for d = 100. Plot of CDF of their pairwise dot products



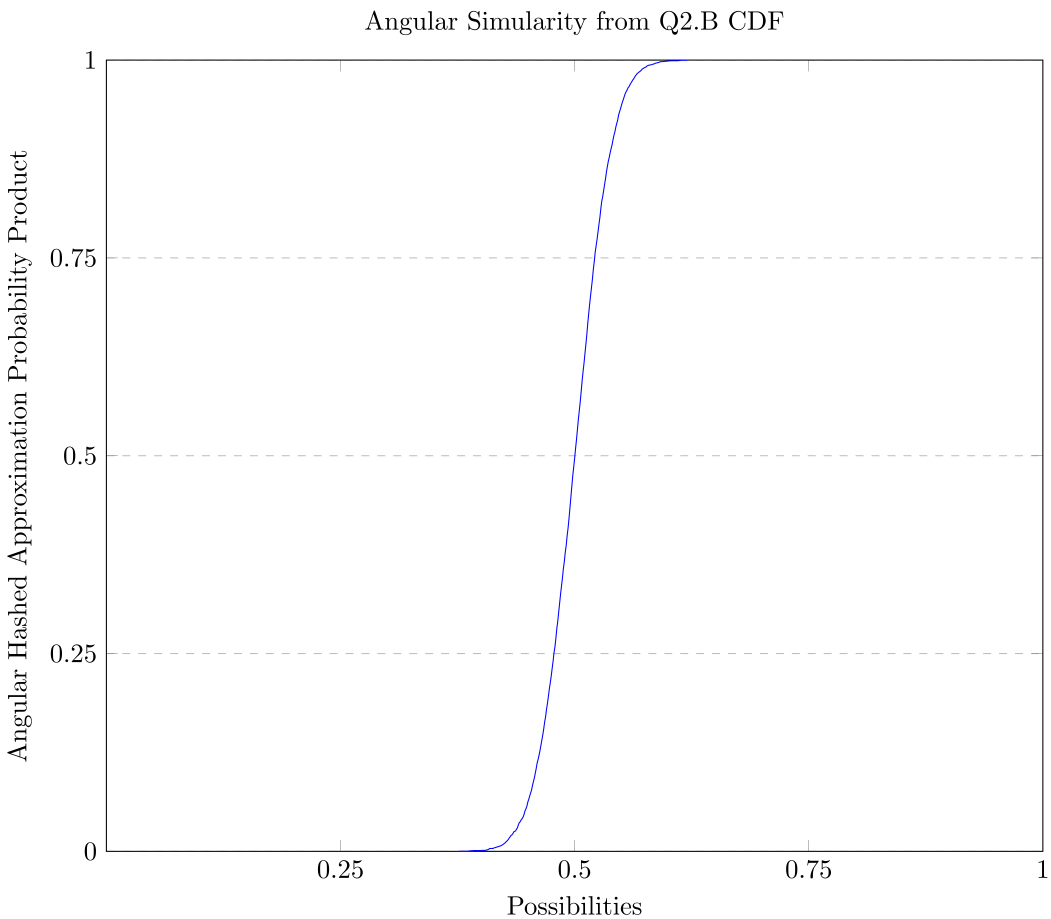
Angular Hashed Approximation

Consider the n = 500 data points in Rd for d = 100 in data set R, given at the top. We will use the angular similarity, between two vectors a, b ϵ R^d: $S_{ang}(a, b) = 1 - \frac{1}{\pi} arccos(< \bar{a}, \bar{b} >)$. If a,b are not unit vectors, then we convert them to $\bar{a} = \frac{a}{\|a\|_2}$ and $\bar{b} = \frac{b}{\|b\|_2}$. The definition of sang (a, b) assumes that the input are unit vectors, and it takes a value between 0 and 1, with as usual 1 meaning most similar.

A: (15 points) Compute all pairs of dot products, and plot a cdf of their angular 2 similarities. Report the number with angular similarity more than τ = 0.85.



B: (20 points) Now compute the dot products and angular similarities among t pairs of the t random unit vectors from Q2.B. Again plot the cdf, and report the number with angular similarity above $\tau = 0.85$.



For part A 67299points above angular similarity .85

For part B there are no points for my example