Assignment 2, CMPT 741.

Maria Babaeva

Due date: Nov 17 2017

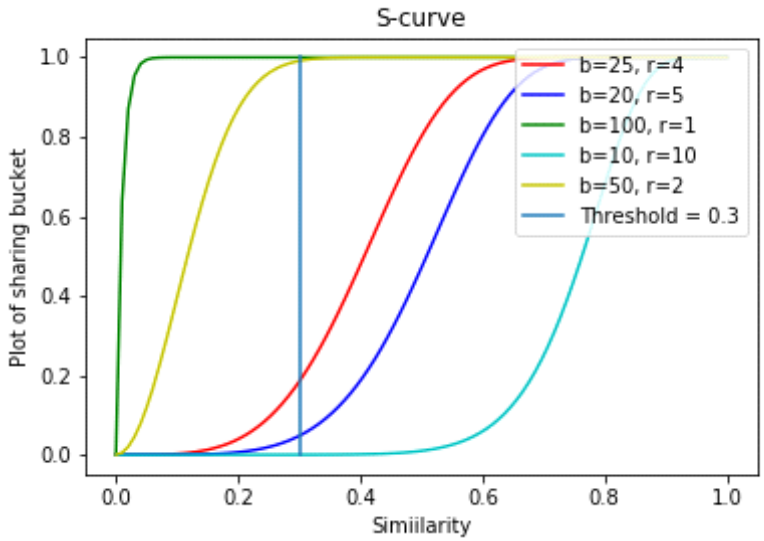
Question 1: report the result in the table below (# refers to the number in the above list)

	3 (b,r)	4(number of pairs)	5(FP,FN)	6(number of remaining pairs after removing FP)	7 (FP,FN)
P=100	(25,4)	35659	33195, 994	1464	293, 6383
P=500	(125,4)	118081	116891, 24	1190	10, 6374

The last step we we needed to compare our candidates with an original matrix gave us a possibility to find a lot of FN pairs which could not be all found after comparing our candidates with a signature matrix.

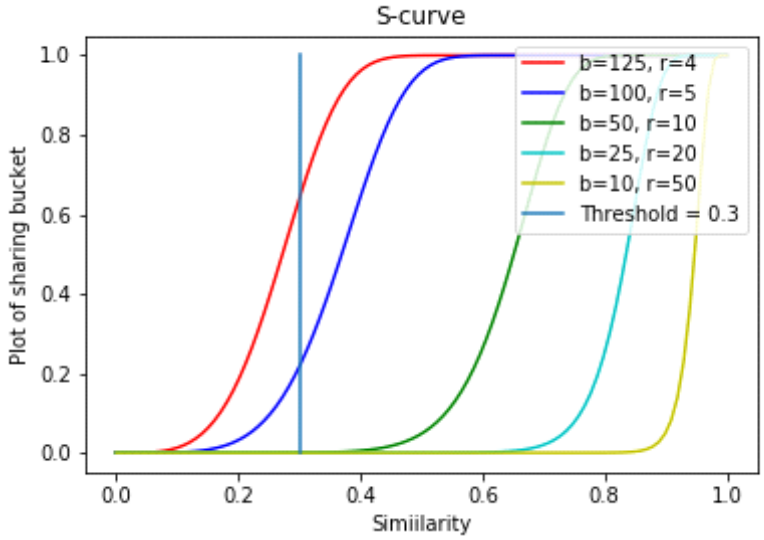
Chosing best b and r for for 100 permut and t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b*r=p. Show the S curve.

Best will be b = 25 and r = 4.



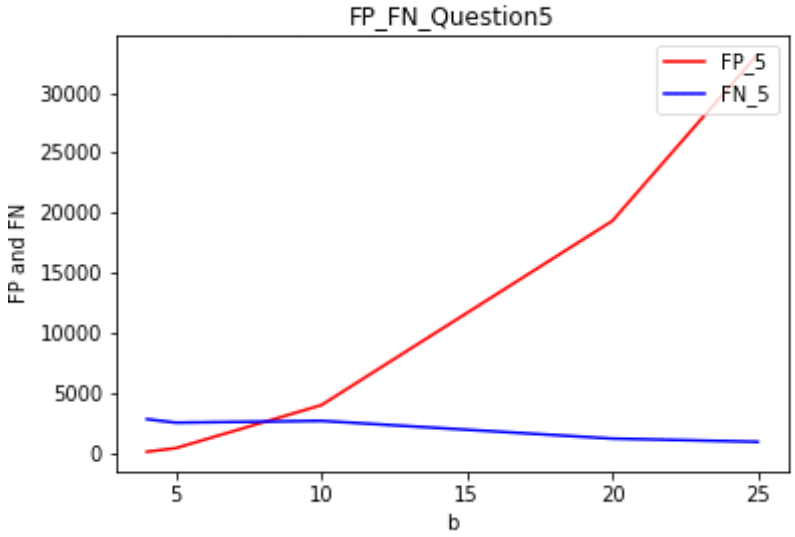
Chosing best b and r for for 500 permut and t=0.3 (i.e., 30%), where b is the number of bands and r is the number of rows per band, and b*r=p. Show the S curve.

Best will be b = 125 and r = 4.

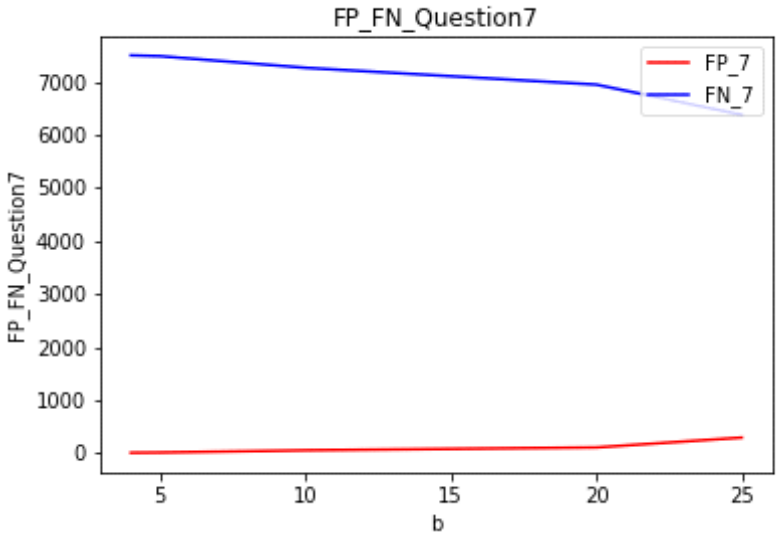


For p=100 permutations:

Question 2: Draw the figure 1 for the FP and FN in 5 for different choices of b



Question 3: Draw the figure 2 similar to figure 1, but for the FP and FN in 7



A short discussion on how the value of b affects FP and FN in 5 and 7.

A proper b and r was calculated using a formula Pr(s) = 1- (1- s^r)^b, where Pr(s) is a probability of sharing the same bucket and s is a similarity of two sets.

A hash function used in 4 is hj = (7* SUM(i,j)(Band) + 157) mod 10000.

The lower b - number of bands the less False Positive (the number of dissimilar signature pairs that were included as candidate pairs) we would have after checking our pairs with pairs from the Signature matrix. The decrease of FP is very significant.

However when we start decreasing the number of bands than the number of similar signature pairs which were not included as our candidate pairs (False Negative) will start increasing.

To find and remove FP first time we would need to calculate Jaccard similarity function to a Signature matrix on choosen candidate pairs and check if their value are less than a treshold = 0.3.

To find and remove FN first time we would need to calculate Jaccard similarity function to our original matrix on (all possible similar pairs from a Signature matrix not considered as candidate pairs) and check if their values are more than a treshold = 0.3.

In the last step we removed all FP and repeat step 5 applying the same logic. The result was that have found a lot of FN pairs which we could not find after comparing our candidates with just a signature matrix.