Stability of Measure Comparisons for avg-idf on msmarco-subsample (p=0.9) Count Delta 0.5 500 1000 1500 2000 0.4 2500 Swap Probability 5.0 0.1 0.0 1.0 1.5 0.5 2.0 2.5 3.0 3.5 4.0 Average Performance Delta