Stability of Measure Comparisons for avg-idf on msmarco-subsample (p=0.75) Count Delta 0.5 600 1200 1800 2400 0.4 3000 Swap Probability 5.0 0.1 0.0 0.5 1.0 2.0 1.5 2.5 3.0 3.5 Average Performance Delta