Stability of Measure Comparisons for wig+100 on beir-webis-touche2020 (p=0.75) Count Delta 0.5 500 1000 1500 2000 0.4 2500 Swap Probability • 0.1 0.0 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.25 Average Performance Delta