lasses Confidence CWL2 Attack against BSD at 1.0% Detection Thresho 100% 80% Classification accuracy 60% 40% MagNet BSD with JSD 20% BSD with Confidence BSD w/o Confidence and JSD 0% 20 60 40 80 Confidence in CWL2 attack