IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection

Rohit Das

October 20, 2019

Intrusion Detection Systems (IDS) have been implemented using various Machine Learning models over the years. However, their robustness in the face of an adversary is something that has been lacking research. While current IDSs can detect between normal and malicious traffic, they may fail if a malicious traffic is disguised as normal. This paper aims to evade current IDSs by generating adversarial malicious traffic using a Wasserstein Generative Adversarial Network (WGAN) and hence test their robustness.

The main idea here is that a black-box IDS will be used to classify half of the data set used (NSL-KDD) and the corresponding learning loss shall be recorded. Based on the predicted labels of the black-box IDS, the discriminator of our IDSGAN will learn those labels and mimic the black-box to a certain extent. The predicted labels of the discriminator will then be fed back to the generator, which will accordingly learn and add noise to reduce its own loss of evading detection by the IDS.

The scope of the paper is limited by the dirth of data on attacks of the categories User2Root (U2R) and Remote2Local (R2L) in the data set NSL-KDD. Hence, further data collection in this area shall be a definite improvement for IDSGAN. Future improvements will be headed towards evasion in distributed computing, cloud computing and edge computing. Our research shall be specifically focused on how IDSGAN can be used to evade IDS for an Internet-of-Things (IoT) network. Such a network may be consisting of homogeneous edge-computing devices or heterogeneous devices like mobile phones, printers, etc., which may have their own IDS models built-in or a single distributed IDS model. Given the data for IDSGAN in evading various black-box models individually, we can create a distributed IDSGAN for the devices in the network and train according to the IDS being used. The IDSGAN will be specifically targeting any network-based intrusion detection system (NIDS) present in the network as it is mainly trained on network data, and not on system/host information, on which a host-based intrusion detection system (HIDS) is based.