Detecting malicious behavior in network and endpoint logs is an extremely challenging task: large and complex data sets, highly dynamic innocuous behavior, and intelligent adversaries contribute to the difficulty. Even analyzing the subtle structures in the normal behavior of network entities is a challenging task.

We present a novel technique for learning latent behavioral structures in large amounts network cyber-security event data. Our technique is inspired by the DeepWalk [Perozzi, Al-Rfou, and Skiena 2014] approach and consists of three main steps. First we generate a property graph from network log information, and run random walks on the resulting graph. We then use random walks to create sentences in a synthetic language and use this language to create Word2Vec model. Our language contains a synthetic grammar of network entity nouns corresponding to graph nodes, behavioral verbs corresponding to typed graph edges, and log property adjectives. We then use k-means clustering on the generated Word2Vec model to find and analyse latent behavioral structures such as behavioral-based clusters of network users and identification of outlying behaviors. We have implemented our technique in Apache Spark and applied it to the public Los Alamos National Laboratory network cyber dataset containing login, network flow, and endpoint process data.

Joint work with Christopher McCubbin (Sqrrl Data, Inc.).