**Slide 1**

Good Morning/Afternoon, I am 2d Lt Newlin and today I will be discussing Evaluation Metrics for Synthetic Data Generation with Machine Learning

**Slide 2**

Here are the topics that I will be covering.

**Slide 3**

Intrusion Detection Systems are a critical aspect of network defense in modern networks. However, to be effective, they need to be trained before being deployed and this can require a large amount of data.

This leads us to the problem. There are currently very few datasets readily available for training NIDS. One reason for this is that the majority of the available datasets are too old, for example the KDD-CUP 1999 dataset which was released in 1999. Due to its age, the traffic it contains doesn’t reflect the types of traffic that we observe today.

Another issue is that some datasets are not truly reflective of the traffic we see today. The Cyber Defense Exercise 2009 dataset contains a much larger amount of malicious traffic than is normally observed in day to day networks.

Finally, another reason for the lack of data is the privacy concerns that are produced by utilizing real data. Often, data that has come from a real source is so heavily anonymized to the point that it is no longer useful.

**Slide 4**

One solution to this lack of data is to synthetically generate data that models real data. This technique is successful when dealing with numbered values. However, it is harder to apply it to generating network data because network data falls into the category of semi-structured data and it is discrete. These properties complicate the synthetic generation process.

One note here is that for the generation process we need to focus on the semantics. We can generate syntactically correct data, but it is much harder to capture the semantic properties of real data.

Thus, the goal here is to intake some real data and then output a synthetic dataset using machine learning.

This leads us to the broad research question that I will be addressing. When generating synthetic data is great, but the question we must ask along the way is “how do we know how good (or real) our synthetic data is?” This is the focus of my research.

**Slide 5**

For some background, the synthetic data generation architecture that I will be dealing with in my research is called the Generative Adversarial Network.

This architecture is made up of a Generator and a Discriminator. The generator creates a fake sample based on the real sample and passes that to the discriminator. The discriminator then must classify the sample as real or fake and this feedback is then passed to the generator.

This game continues until the discriminator cannot discriminate the fake samples from the real ones.

GAN is a deep learning architecture, and the generator and discriminator are usually some form of Neural Network

So far the architecture is successful with images and the bottom figure displays a bunch of different generated images from a GAN and at a glance they look pretty real, although on closer inspection you can see that they aren’t real.

Currently though, the GAN architecture struggles with generating text based data because the discrete nature of text prevents the gradients from being propagated from the discriminator to the generator.

**Slide 6**

There has been almost continual improvement on the GAN architecture since its introduction in 2014. The two most notable are SeqGAN and TreeGAN.

SeqGAN generates sequences of text and additionally introduces some reinforcement learning into the feedback process to address the issues with discrete data.

TreeGAN combines SeqGAN with a tree-based architecture and a context free grammar to produce text data that is syntax aware.

On the evaluation criteria side, there is some work being done on finding different quantitative metrics that can be used to measure how real the outputs from a GAN are. Currently, all of this work deals images and there has been little work on the text side.

**Slide 7**

So what I will be doing is looking into the evaluation metrics for GANs on the text based side of the house. Specifically, I will be feeding network logs into a GAN architecture and then evaluating the outputs with different metrics to see which ones perform the best. The motivation for this is that currently the main metric for quality of output is still human evaluation so we need quantitative not qualitative metrics so that we can remove the human evaluation piece.

From my exploration of the literature, it looks like my work may incorporate some elements from Natural Language Processing.

The figure on the screen is an example of a real SQL query and a synthetic SQL query from TreeGAN. My reasoning for showing this is to demonstrate that we need a quantitative way to evaluate how “real” this synthetically generated SQL query is.

Ultimately the hope is that by developing better metrics we can develop better synthetic data.

**Slide 8**

These are the topics I discussed today

**Slide 9**

This concludes my brief, thank you for your time and are there any questions?