# 3 Jun 19

MEMORANDUM FOR: AFIT/ENG

ATTENTION: LT COL DEYOUNG

FROM: 2d Lt Marvin Newlin (GCO-20M)

SUBJECT: Thesis Prospectus: *Evaluation Metrics for Synthetic Text Generation*

1. Intrusion Detection Systems (IDS) are a critical aspect of cybersecurity in modern networks. Training IDS requires large amounts of data in the form of log files or packet captures, both of which are forms of text data. Due to privacy concerns, there is a lack of real data available to train IDS. The ability to synthetically generate text that is semantically equivalent to real text will allow for the development of synthetic datasets that can be used to train IDS. The goal of this research is to explore and possibly develop metrics that can allow for generation of synthetic text data that is semantically equivalent to real text data.

2. Currently, synthetically generating text that is syntactically correct is possible. However, there exists a lack of quantitative metrics for evaluating the semantic quality of synthetically generated text. Quantitative methods are necessary due the large amounts of data that are required for real-life applications of synthetic data. Currently, the idea of the Generative Adversarial Network (GAN) is being utilized to successfully produce synthetic images and work is being done to develop quantitative metrics for evaluating the quality of generated images. This successful generation relies on the fact that images can be represented in a continuous space, however, text data lives in a discrete space, so the current methods for GANs cannot be directly applied to text data. Analyzing existing metrics for quantitative evaluation of generated text is the goal of this research.

3. This research will evaluate current available metrics for synthetic text evaluation in the context of GANs and other generative methods. Additionally, this research aims to improve upon existing metrics and possibly develop new metrics that can accurately capture the semantic properties of real text data, in order to produce more realistic text data. Currently, the only outside resources necessary for this research is the DoD High Performance Computing (HPC) system, to which we already have access.

4. The results of this research will be a performance comparison of existing metrics and how well they model real text data. Additionally, any improvements or new metrics developed will be compared to existing solutions to show improvement over the current methods.

5. The results of this work will allow for further development of synthetic text that more accurately models real text. As a result, the synthetic generation of datasets will be possible. This will allow for the ability to produce data that can model any number of possible problems that current IDS or other cybersecurity assets that require datasets need.

6. Proposed thesis committee:

a. Lt Col Mark DeYoung, Chair / Thesis advisor \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(signature)*

b. Dr. Lawrence Merkle?, Committee member \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(signature)*

c. Dr. Clark Taylor?, Committee member \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(signature)*

7. Sponsor: N/A

8. Listed below are the courses that augment my research:

a. CSCE 586: Algorithms

b. CSCE 523: Artificial Intelligence

c. CSCE 623: Introduction to Statistical Machine Learning

d. CSCE 554: Fundamentals of Experiment Design and Performance Analysis

MARVIN W. NEWLIN, 2d Lt, USAF

GCO-20M

1st Ind, AFIT/ENG

MEMORANDUM FOR AFIT/ENG

I acknowledge receipt of the above thesis prospectus and thesis committee. This prospectus will be maintained in the department files for students graduating between Sept 2019 and Jun 2020. The thesis should be prepared in accordance with the AFIT Thesis Guide. Good luck!

MARK E. DEYOUNG, Lt Col, USAF, PhD

Chief, Computer Science Division

Department of Electrical and Computer Engineering