# *Biomedical Informatics Grand Rounds*

**Wednesday, Sept. 15, 2021 3:00 pm – 4:00 pm**

# Detection of Trojan Attacks to Deep Neural Networks – A Topological Perspective

*Chao Chen, PhD*
*Assistant Professor,*
*Department of Biomedical Informatics, Computer Science and Applied Mathematics & Statistics,*
*Stony Brook University*
*Stony Brook, NY*

**Remote Access**
**Join Zoom Meeting** https://stonybrook.zoom.us/j/95617197636?pwd=KytzZ2pVRG9SZGpKZUtpNXJISjNjZz09
Meeting ID: 956 1719 7636 Passcode: 924293

**Bio:** Dr. Chao Chen is an Assistant Professor at Stony Brook University. His research interests span topological data analysis (TDA), machine learning, and biomedical image analysis. He develops principled learning methods inspired by the theory from TDA, such as persistent homology and discrete Morse theory. These methods address problems in biomedical image analysis, robust machine learning, and graph neural networks from a unique topological view. His research results have been published in major machine learning, computer vision, and medical image analysis conferences. He is serving as an area chair for MICCAI, AAAI, CVPR, and NeurIPS.

Dr. Chen received his Ph.D. in Computer Science from Rensselaer Polytechnic Institute, a Master's degree from the National University of Singapore, and a B.Sci. degree from Peking University, China. He joined the Biomedical Informatics Department in 2018. Before coming to Stony Brook, he was a faculty member at the City University of New York.

**Abstract:** Deep neural networks are known to have security issues. One particular threat is the Trojan attack. It occurs when the attackers stealthily manipulate the model's behavior through Trojaned training samples, i.e., samples with special triggers injected and labels altered. Identifying a Trojaned model at deployment is challenging, due to limited access to the training data. We propose different approaches to identify Trojaned neural networks by (1) inspecting high-order topological features of the neuron interactions and (2) reverse-engineering the injected triggers using a topological loss. These approaches take different angles and reveal insights into the behavior of neural networks when their strong memorialization power is exploited maliciously. We will also briefly review other works such as how to train a robust model with label noise, and how to improve the robustness of graph neural networks against structural attacks.

**Educational Objects**: Upon completion, participants should be able to:

- Understand Trojan attacks to neural networks
- Learn the label noise issue of neural network training
- Learn different methods to identify attacked models

**Disclosure Statement**: In compliance with the ACCME Standards for Commercial Support, everyone who is in a position to control the content of an educational activity provided by the School of Medicine is expected to disclose to the audience any relevant financial relationships with any commercial interest that relates to the content of his/her presentation.
The faculty: *Chao Chen, Ph.D.*, the planners; and the CME provider have no relevant financial relationship with a commercial interest (defined as any entity producing, marketing, re-selling, or distributing health care goods or services consumed by, or used on, patients), that relates to the content that will be discussed in the educational activity.

**Continuing Medical Education Credits**: The School of Medicine, the State University of New York at Stony Brook, is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The School of Medicine, the State University of New York at Stony Brook designates this live activity for a maximum of **1** ***AMA PRA Category 1 Credits™.*** Physicians should only claim credit commensurate with the extent of their participation in the activity.