Introduction: With the spectacular advances in Deep Learning over the past decade, humanity has gained new tools to solve previously intractable problems. These include superhuman machine performance in strategy games such as Go and StarCraft, and simplifying large, macroscopic problems in math and physics like discovering new algorithms for matrix multiplication. At the intersection physics and deep learning there has been cutting edge research to learn how to embed *network* (graph) data into a deep learning framework [1] [2] – potentially allowing deep learning to explore network data and discover network weaknesses. Indeed, real-world networks such as power grids are fundamentally vulnerable to attacks on a small number of key components (nodes), whose removal can cause the system as a whole to collapse. Though identifying these "Achilles Heels" of a given network is a computationally intractable (NP-Hard) problem [3], deep learning might effectively surmount this obstacle, providing new destabilizing tools that malicious agents could use to attack real-world infrastructure networks. To our knowledge, little research has sought to understand the limits to deep learning's capacity to efficiently destroy networks, and counterstrategies that might mitigate this capacity.

Objectives: To understand the limits of deep learning's ability to dismantle networks by removing key nodes. And should no such limits exist, to develop AI-based counterstrategies to in a "fight fire with fire" approach.

Hypothesis: I hypothesize that deep learning will have the capacity to discover key network features at decreasing rates as more information about the structure of the network is hidden from the deep learning agent (an attacker). Further, I hypothesize this trend will hold for networks of varying sizes and degrees of interconnection. I suggest that the most efficient concealment strategy can be discovered by a *second* deep learning agent (an opposing defender), tasked with learning to undermine the first agent through tandem learning strategies – allowing for the creation of strategies to defend real world networks and infrastructure.

Methods: I will create a framework which maps the real-world problem of network attack and defense to a two-player strategy game that can be solved with Deep Graph Learning. Specifically, my framework will replicate network attack through the network dismantling effects of node destruction – a deterministic process with sequential outcomes resembling percolation [3] [4]. For the attacker, the game will be defined by taking actions on a set of nodes which will break down the network, ending the process when a specified portion of the network is fragmented from the rest of the network. For the defender, the network learning problem will be defined by taking actions on a set of links until a certain fraction of the network is concealed. As such, the attacker will attempt to make its decision based only the information left after the defender has obscured part of the network.

The game will be played on networks of varying size and degrees of interconnection. To this end, I will use network embedding to capture important network features, transforming the structure of the data into a fixed dimensional tensor suitable for deep reinforcement learning [1]. Through expanding this approach to allow for link selection, not only node selection, I enable a deep learning agent (the defender) to learn to conceal network links, reducing the effectiveness of the node selecting deep learning (the attacker). I will reward each player based effectively they shorten or lengthen the number of moves taken by the attacker – incentivize each of their defined objectives.

Anticipated Results: Using this approach, I will determine whether the deep learning of weak points of a network can be thwarted by strategic concealment of key network information, and what the optimal such defensive strategy might be. If an attacker's ability to quickly dismantle networks cannot be reduced in a meaningful way, this might reveal a fundamental weakness of infrastructure and other-real world networks to malicious attacks by a deep-learning equipped adversary [5] [6] [7].

Conclusion and Significance: The deliverable from this research will be defensive strategies which thwart malicious deep learnings seeking to destroy networks.

References

[1] H. Dai, E. B. Khalil, Y. Zhang, B. Dilkina and L. Song, "Learning Combinatorial Optimization Algorithms over Graphs," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[2] C. Fan, L. Zeng, Y. Sun and Y.-Y. Liu, "Finding key players in complex networks through deep reinforcement learning," *Nature Machine Intelligence,* vol. 2, pp. 317-324, 2020.

[3] H. Bennett, D. Reichman and I. Shinkar, "On Percolation and N P-Hardness," *Random structures & algorithms,* vol. 54, no. 2, pp. 228-257, 2019.

[4] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature,* vol. 514, pp. 65-68, 2015.

[5] X.-L. Ren, N. Gleinig, D. Helbing and N. Antulov-Fantulin, "Generalized network dismantling," *PNAS,* vol. 116, no. 14, pp. 6554-6559, 2018.

[6] M. Grassia, M. De Domenico and G. Mangioni, "Machine learning dismantling and early-warning," *Nature Communications,* vol. 12, no. 1, 2021.

[7] R. Albert, H. Jeong and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature,* vol. 406, p. 378–382, 2000.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves and I. Antonoglou, "Playing Atari with Deep Reinforcement Learning," *ArXiv,* vol. abs/1312.5602, 2013.