Research proposal

Topic: **Adversarial Social Detection based on Graph Neural Networks**

IDEAS Supervisor: **dr Tomasz Michalak** (IDEAS NCBR & University of Warsaw)

Short Description

Social Networks have become a primary media for cybercrimes. For instance, attackers may compromise accounts to diffuse misinformation (e.g., fake news, rumors, hate speeches, etc.) through a social network. Fraudsters may also trick innocent customers into conducting fraudulent transactions over online trading platforms. Meanwhile, on the defense side, defenders (e.g., network administrators) are increasingly employing machine-learning-based tools to detect malicious behaviors. Graph Neural Networks (GNNs) have become the de facto choice of social detection tools due to their superior performance over a wide spectrum of tasks.

In this project, the overall goal is to develop robust and effective GNN-based social detection tools in an adversarial environment. This goal is decomposed into three coherent objectives. First, design more effective GNN-based tools to detect crimes in social networks that could achieve a better detection accuracy as well as a lower false positive rate. Second, from the standpoint of an attacker, investigate effective evasion techniques to bypass the detection of the GNN-based tools. Third, as a defender, enhance the robustness of the GNN-based detection tools to mitigate evasion attacks. Overall, the expected outcomes significantly advance our knowledge in developing trustworthy AI systems in a real-world adversarial environment.

Y. Vorobeychik and M. Kantarcioglu. Adversarial machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 12(3):1–169, 2018