CSCE 676 600: REPORT 2

Name: Rohan Chaudhury

UIN: 432001358

24th March 2022

Paper Title: All You Need Is Low (Rank): Defending Against Adversarial Attacks on Graphs

Paper Citation: Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. 2020. All You Need Is Low (Rank): Defending Against Adversarial Attacks on Graphs. In The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3336191.3371789

1. SUMMARY:

a. Motivation:

In the field of Artificial Intelligence and Machine Learning, adversarial machine learning is a method that tries to deceive machine learning models using false data with the intention of causing a machine learning model to malfunction. Presenting a model with erroneous or misrepresentative data during training, or injecting maliciously prepared data to confuse an already trained model, are both examples of adversarial attacks. According to recent notable research which looked at the effects of adversarial attacks on graph data, it found that graph embedding techniques are susceptible to adversarial attacks as well. The authors in this paper have explored ways to prevent generative adversarial attacks on graph data since according to them, it is highly important to design graph analysis algorithms that can withstand adversarial attacks as graphs are frequently utilized in many areas because of their ability to depict real-world data in a variety of disciplines such as social, citation, and biological networks. The authors investigated the features of the perturbations introduced by the NETTACK model in order to find appropriate countermeasures against them. The authors chose this model because it has proven to be highly effective in misleading the Graph Convolutional Network model and some node classification approaches such as node embeddings. They have also presented a low-rank adversarial attack called LowBlow that can impair the classification accuracy of both tensor-based node embeddings and the Graph Convolutional Network model.

b. Main technical contributions and previous work leading to it:

Previous research in attempts to "vaccinate" a network:

The word "vaccinate" refers to equipping a network with defenses against adversarial attacks. In one research, JPEG compression was employed to "vaccinate" a deep neural network as they theorized that adversarial attacks create distortions to the high-frequency band in images that are visually undetectable, and so JPEG compression can effectively eradicate them. Bhagoji et al. suggested a defense method that uses Principal Component Analysis for a reduction in dimensionality in another study. Another study uses Singular Value Decomposition and low-rank approximation of the adjacency matrix of individual users in Amazon and Twitter to identify strange activities when

dealing with physical adversaries, with the primary objective being the detection of dense block-like behavior that points to fraud such as fake followers in Twitter. There are two types of adversarial attacks, evasion and poisoning attacks. In evasion attacks, adversarial data is updated at test time to get around the result and in poisoning attacks which is the type being investigated in this paper, the training data is perturbed and the classification model is retrained using this perturbed data.

Concepts used by the authors:

Singular value Decomposition:

SVD is used to calculate the best rank-r approximations of a given matrix A. The SVD of a matrix A is computed in the following way:

$$A = U\Sigma V^T$$

Where U and V are orthogonal matrices and Σ is a non-negative diagonal matrix. The rank-r approximation (A_r) of the matrix A can be calculated as follows:

$$A_r = U_r \Sigma_r V_r^T = \sum_{i=1}^{r} u_i \sigma_i v_i^T$$

Here:

U_r and V_r are two matrices that contain the top r singular vectors and Σ_r is a diagonal matrix that contains r singular values. So, using SVD we can derive the rank-r approximation of any real-valued matrix A.

Tensors:

A tensor is a multidimensional matrix. The number of indices necessary to index a tensor is its order. Three-mode tensors are the subject of this study.

Methods used and investigations conducted by the authors:

Investigation of the effects of NETTACK on graph data:

The authors used the NETTACK model (proposed by Zügner et al.) to create adversarial attacks to perturb Graph Convolutional Networks (GCN). NETTACK creates undetectable perturbations by placing various constraints on the attacks to ensure that the graph structure and node attributes are preserved. Provided a Graph G = (A, X), where A is the undirected adjacency matrix and X is the feature matrix, any changes to matrix A are referred to as structure perturbations, and any changes to X are referred to as feature perturbations. Attacks that maintain the graph's degree distribution are deemed undetectable when it comes to generating imperceptible structure perturbations and the co-occurrence of the features is taken into account to produce imperceptible feature perturbations. The authors investigated the adjacency and feature matrices before and after the attack to empirically test their intuition and understand the features of the perturbations caused by the NETTACK model. To show the differences, the authors displayed the singular values of matrices for the original and the attacked graphs. **Figure 1** shown below depicts the singular values of the adjacency matrix before and after one attack on the target node on a semi-logarithmic scale. From the figure, we can see that at higher ranks only, the original and attacked matrices' singular values have a significant difference. Then, singular values of adjacency and feature matrices for several

attacks (more than one) were investigated by the authors, and they discovered that singular values are fairly similar at lower ranks but differ at higher ranks. They use this idea to provide two low-rank methods that can successfully defend against NETTACK.

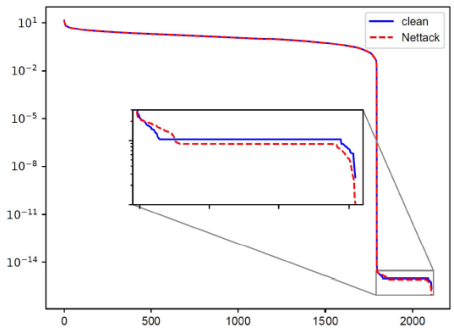

Figure 1: The singular values of the adjacency matrix before and after one attack on the target node on a semi-logarithmic scale

Low-rank solutions proposed by the authors to defend against attacks:

The authors constructed the low-rank estimation of the feature and adjacency matrices by extracting them from their SVD decomposition in order to eliminate the high-rank disturbances induced by NETTACK. The GCN was then retrained using the low-rank approximation matrices. The authors demonstrated that, with the correct choice of r, the rank-r estimation of the attacked graph can increase GCN's performance and get it closer to that of GCN trained on the original graph. Essentially, computing the graph's SVD exposes the attack's spectrum as well as the graph's healthy regions. The authors are then keeping a truncated SVD that only contains the top-k singular values for the graph and then rebuilding the graph from it as they are assuming high-rankness of NETTACK on the basis of their experiments. The vaccinated graph is the result of such truncation.

t-PINE: Tensor-Based Node Embeddings with High Robustness

Al-Sayouri et al. in their paper had described a tensor-based node embedding approach that uses the tensor's CP decomposition (known as CANDECOMP/PARAFAC) to capture node-to-node relations using low-dimensional latent components. In their paper, they had evaluated the performance of t-PINE with node embeddings, but they didn't test its resilience in an adversarial situation. The authors of this paper deemed t-PINE to be a suitable choice to defend against high-rank perturbations created by NETTACK because of its intrinsic low-rank character, and so they used this approach in their experiments.

LowBlow - a proposed attack of Low-Rank:

Changes in high-rank singular values are caused by NETTACK disruptions, which may be countered using a low-rank graph approximation. Low-rank attacks, on the other hand, may still be able to affect the graph data. To demonstrate whether low-rank attacks have an impact on graph data, the authors adjusted the NETTACK perturbations to produce low-rank attacks that can impede both GCN and t-PINE's performance.

The authors also demonstrated that modifying LowBlow to only have the edges that maintain the degree distribution of the graph makes it a high-rank attack akin to NETTACK.

c. **Interesting experimental result/theoretical finding:**

The authors have utilized the largest connected component of Cora-ML, CiteSeer, and PoliticalBlogs graph datasets for their experiments. To evaluate the effectiveness of their proposed defensive measure the authors computed the effectiveness using different values of rank r.

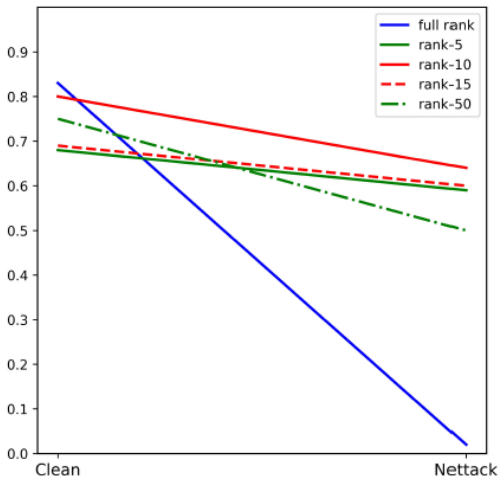

Figure 2: Percentage of target nodes that were correctly classified for the NETTACK perturbed graph and the original (clean) graph for various rank approximations of the feature and adjacency matrices

The graph in **Figure 2** above shows the percentage of target nodes that were correctly classified for the NETTACK perturbed graph and original (clean) graph for various rank approximations (the values of rank r under test being 5, 10, 15, and 50) of the feature and adjacency matrices. As we can see from **Figure 2**, with full-rank attacked matrices, the proportion of target nodes properly classified after the attack reduces dramatically. But the fraction of properly classified nodes on the attacked graph is close to that of the original graph when employing the low-rank SVD approximation, as can be seen from **Figure 2.** From the graph, we can say that by utilizing rank-10 approximation of the adjacency and feature matrices (i.e., using only ten singular values/vectors) the impact of NETTACK can be considerably reduced (rank-10 approximation provides the best results). A rank-10 approximation can provide a reliable approximation of the graph structure and characteristics, as well as can "vaccinate" Graph Convolutional Networks against adversarial attacks.

The authors also tested the robustness of t-PINE against NETTACK and found that it is quite resilient for lower embedding dimensions, but as the dimension grows larger, the resilience of the embedding also declines. Also, the authors found that LowBlow drastically reduces the performance of both t-PINE and GCN but GCN is affected more by LowBlow than t-PINE. The authors also tested their proposed defense method against LowBlow by taking a rank-10 approximation of the graph in order to "vaccinate" it. From this experiment, they observed that LowBlow is harder to resist than NETTACK because of its low rank. Also from the results, it can be concluded that the "vaccination" strategy proposed by the authors performs better on NETTACK than LowBlow since LowBlow is inherently a low-rank attack.

2. MY THOUGHTS:

a. Pros and challenges they address:

i. Adversarial attacks are a major concern in the field of Artificial Intelligence and Machine Learning. Previous work addresses the issues of adversarial attacks on

machine learning models to design defense strategies that are resilient to said attacks, however, there is relatively less research conducted in the area of effects and mitigations against adversarial attacks on graph data. Because of the non-linear structure of Graph convolutional networks, they have demonstrated tremendous success in tasks related to node classification but despite their success, they are vulnerable to minor disturbances. So, this research is aimed to deliver a detailed study on the effects of adversarial attacks on graphs and investigate some effective defense mechanisms against them. Since graphs are utilized in many different areas such as social, citation, and biological networks, this research addresses a vast domain of concerns.

ii. The authors investigated the properties of NETTACK perturbations and successfully demonstrated that they cause significant changes in the graph's high-rank spectrum, which correlates to low singular values.

iii. The authors show that when the graph's low-rank approximation is utilized, the GCN model can considerably withstand the attacks, based on the concept that the Nettack perturbations are high-rank.

iv. The authors demonstrated that tensor-based node embeddings that approximate a low-rank representation of the graph are particularly resistant to adversarial attacks by showing the impact of the NETTACK model generated adversarial high-rank attacks on a tensor-based node embedding approach.

v. The authors developed the LowBlow attack, which alters NETTACK perturbations to influence low-rank graph components. As a result, the novel low-rank attack can mislead both tensor-based embeddings and the GCN.

vi. The authors demonstrate that changing a low-rank attack to retain the graph's degree distribution transforms it into a high-rank attack.

b. Limitations:

i. The paper investigates only the poisoning type of adversarial attacks to build defenses against them with no investigation regarding the evasion type of adversarial attacks on graph data.

ii. Although the research claims that the singular values of the adjacency and feature matrices before and after perturbations by NETTACK are fairly similar at lower ranks but differ at higher ranks it can be possible that such behavior is not observed in every possible scenario and counterexamples of such an assumption may be present. More research is required to identify possible pitfalls of such an assumption and to find possible ways to defend against adversarial attacks without such assumptions.

iii. Only one type of tensor-based node embedding approach was evaluated in this study. A comparative study with other tensor-based node embedding approaches is required for better evaluation of the task at hand since other tensor-based node embedding approaches should, in theory, provide similar low-rank approximations of the graph that according to the authors should be able to resist NETTACK perturbations.

iv. The research only deals with adversarial attacks generated by NETTACK and LowBlow. However, it is not determined how well the low-rank approximation method would hold up against real-world adversarial attacks on graph data.

3. FOLLOW UP/NEXT STEPS:

a. Poisoning attacks are rare as compared to the evasion type of adversarial attacks. This paper only investigates and tries to build defenses against the poisoning type of attacks on graph data. Future work can be undertaken to build defenses against the evasion type of attacks on graph data.
b. Research can be undertaken to determine ways to defend against adversarial attacks without the assumption that singular values of the adjacency and feature matrices before and after perturbations by an adversarial attack are fairly similar at lower ranks but differ only at higher ranks.
c. Research can be undertaken by utilizing several types of tensor-based node embedding approaches and then observing whether the same results hold for them or not. Other tensor-based node embedding approaches should, in theory, provide similar low-rank approximations of the graph which can be utilized to see if they provide resistance against NETTACK perturbations.
d. This research deals with adversarial attacks generated by NETTACK and LowBlow. Further research can be undertaken to test out adversarial attacks generated by various other approaches and if possible, against real-world adversarial attacks.

4. REFERENCES:

a. Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. 2020. All You Need Is Low (Rank): Defending Against Adversarial Attacks on Graphs. Proceedings of the 13th International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, 169–177. DOI:https://doi.org/10.1145/3336191.3371789
b. https://www.youtube.com/watch?v=CpD9XlTu3ys
c. https://www.youtube.com/watch?v=PFDu9oVAE-g
d. https://www.youtube.com/watch?v=KTKAp9Q3yWg