Dai et al. [1] classified graph adversarial defense algorithms into adversarial training, provable robustness, and graph purification.

The main idea of adversarial training is to inject adversarial samples into training set so that the trained model can correctly classify future adversarial samples. Xu et al. [2] explored adversarial training on GNNs for the first time. Deng et al. [3] and Feng et al. [4] applied virtual graph adversarial training to further promote the smoothness of model predictions on labeled and unlabeled nodes.

The main idea of provable robustness is to know how worst-case adversarial at-tacks affect the model and maximizes the worst-case margin directly during training to encourage the model to learn more robust weights. Zügner et al. [5] ﬁrst studied the provable robustness to perturbations of node features. Zügner et al. [6] also proposed a branch-and-bound algorithm that can obtain a tight bound on the global optimum for topology attack demonstration. A stochastic smoothing technique was applied in Wang et al. [7] to provide provable guarantees for any GNN.

The main idea of graph purification is that, first, we denoise the graph with heuristics for graph-intrinsic properties or attack behavior; then, we can train the GNN model on the denoised graph to give correct predictions that are not affected by the poisoned attacker. Wu et al. [8] proposed GCNJaccard to resist adversarial attacks by eliminating edges connecting nodes with low Jaccard similarity. Entezari et al. [9] used a low-rank approximation of the truncated SVD to denoise the graph based on the observation of high-rank attacks to resist poisoning attacks. Jin et al. [10] proposed to learn a clean adjacency matrix with low rank, sparse, and smooth features.

1. Dai E, Zhao T, Zhu H, et al. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability[J]. arXiv preprint arXiv:2204.08570, 2022.
2. Xu K, Chen H, Liu S, et al. Topology attack and defense for graph neural networks: An optimization perspective[C]//International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, 2019.
3. Deng Z, Dong Y, Zhu J. Batch virtual adversarial training for graph convolutional networks[J]. arXiv preprint arXiv:1902.09192, 2019.
4. Feng F, He X, Tang J, et al. Graph adversarial training: Dynamically regularizing based on graph structure[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(6): 2493-2504.
5. Zügner D, Günnemann S. Certifiable robustness and robust training for graph convolutional networks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 246-256.
6. Zügner D, Günnemann S. Certifiable robustness of graph convolutional networks under structure perturbations[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 1656-1665.
7. Wang B, Jia J, Cao X, et al. Certified robustness of graph neural networks against adversarial structural perturbation[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 1645-1653.
8. Wu H, Wang C, Tyshetskiy Y, et al. Adversarial examples for graph data: deep insights into attack and defense[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 4816-4823.
9. Entezari N, Al-Sayouri S A, Darvishzadeh A, et al. All you need is low (rank) defending against adversarial attacks on graphs[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 169-177.
10. Jin W, Ma Y, Liu X, et al. Graph structure learning for robust graph neural networks[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 66-74.