Adversarial Edge Perturbation Framework in Graph-based Retrieval

Abstract. Graph-based retrieval systems leverage structural and semantic relationships among entities to enable context-aware search. However, their reliance on graph structure renders them vulnerable to adversarial perturbations of graph nodes and edges that distort embeddings and alter ranking outcomes. This paper introduces an approach for *adversarial edge removal* for targeting graph-based retrieval. We first establish that conventional structural heuristics, such as centrality degrees or PageRank, are non-deterministic predictors of rank degradation under edge perturbations, as embedding shifts depend on multi-hop spectral interactions rather than first-order topology. Building on this, we propose a learning-based estimator that models the mapping between local edge characteristics and their induced ranking distortion. The estimator, trained on observed perturbation–response pairs, enables efficient identification of high-impact edges within a constrained budget, operating in a black-box setting. Experiments on benchmark graph retrieval datasets demonstrate that the proposed framework achieves stronger and more efficient rank demotion than state-of-the-art baselines.

1 Introduction

Graph-based retrieval systems form a core component of contemporary information access pipelines, enabling structured search over entities connected through semantic or relational links. These systems integrate textual content with topological context, allowing relevance estimation to account not only for term-level similarity but also for relational proximity within the graph [1,13]. The effectiveness of such systems, however, is naturally tied to the stability of the underlying graph structure. Minor perturbations in connectivity can propagate through embedding functions, altering the geometric representation of nodes in latent space and consequently reshaping retrieval rankings. These perturbations may arise unintentionally from *data noise* or deliberately through *adversarial manipulation*, both of which threaten the reliability of graph-based retrieval. A rigorous understanding of how structural changes translate into ranking shifts is, therefore, fundamental to the design of robust graph-based retrieval systems.

Adversarial attacks on graphs can be classified based on *timing* and *attacker knowledge* [16]. With respect to timing, *poisoning attacks* modify the graph before model training, whereas *evasion attacks* occur after the model has been trained. In terms of knowledge, *white-box* attackers possess full access to model parameters and gradients, *gray-box* attackers operate with partial information, and *black-box* attackers have no access, relying solely on observable structural or embedding cues. This work focuses on *black-box poisoning attacks* in which an adversary removes a small number of edges prior to the embedding process. This setting is both realistic and consequential for retrieval systems, since subtle manipulations to the graph structure before training can distort node representations and thereby alter the ranking of target nodes during retrieval.

Existing studies on structural attacks have primarily explored two methodological directions. Gradient-based approaches leverage model gradients to optimize perturbations [17,3], achieving high precision but requiring full access to model parameters,

which is impractical in restricted or proprietary environments. Heuristic methods, in contrast, select edges based on structural indicators such as degree, homophily, or PageRank [4,11,10]. These methods are computationally efficient but rest on the assumption that structural importance correlates with adversarial sensitivity. This assumption does not necessarily hold in retrieval contexts, where node embeddings are produced through non-linear message passing and spectral interactions within the graph Laplacian rather than first-order node statistics. Hence, edges that appear structurally insignificant can have disproportionate influence on embeddings and, consequently, on ranking outcomes.

We theoretically show that no monotonic function of local structural metrics (e.g., degree or PageRank) can predict the rank degradation of a target node after edge removal. Because embedding perturbations depend on multi-hop propagation and spectral coupling, an edge's impact on retrieval relevance is inherently a non-linear function of the graph topology and embedding process. This key limitation of heuristic-based attacks motivates a learning-based approach that directly models the relationship between structural perturbations and retrieval degradation. We propose a data-driven approach for adversarial *edge removal* that learns to estimate the effect of each edge on retrieval rankings. Rather than relying on gradient access or retraining, the framework trains a neural estimator to approximate the mapping between local edge representations and their induced ranking degradation. Once trained, the model efficiently identifies high-impact edges within a constrained perturbation budget, enabling targeted rank demotion. The contributions of this work are threefold: **(1)** We provide a formal analysis showing that traditional structural heuristics are non-deterministic predictors of adversarial sensitivity in graph-based retrieval. **(2)** We introduce a learning-based framework that estimates the influence of individual edges on ranking outcomes through data-driven approximation rather than structural assumptions. **(3)** We demonstrate, through experiments on real-world graph retrieval benchmarks, that the proposed method achieves stronger and more efficient rank demotion than both heuristic and gradient-based baselines.

2 Methodology

Task Definition. Let $G = (V, E)$ denote an undirected and unweighted graph with adjacency matrix $A \in \{0,1\}^{|V|\times|V|}$, where V and E represent the node and edge sets, respectively. Each node $v \in V$ is associated with an embedding vector $\mathbf{z}_v \in \mathbb{R}^K$ obtained from a graph representation learning algorithm that encodes both structural and semantic information. Given a query representation $\mathbf{q} \in \mathbb{R}^K$, nodes are ranked according to their similarity to $\mathbf{q}$, typically measured by cosine similarity. Let $r(v_t)$ denote the ranking position of a target node v_t in this ordering. The attacker's goal is to identify and remove a small subset of edges whose deletion causes the greatest deterioration in the ranking position of v_t. Formally, the attack seeks $e^* = \arg\max_{e\in E} \Delta r(v_t, e), \quad \text{where} \quad \Delta r(v_t, e) = r'(v_t) - r(v_t)$, subject to a perturbation budget B that limits the number of modified edges, i.e., $|E'| - |E| \leq B$, where E' denotes the altered edge set. In essence, the attack aims to discover structurally critical edges whose removal most strongly disrupts the learned embedding relationships.

2.1 Limitations of Structural Heuristics

Structural metrics such as node degree and PageRank are commonly used to estimate edge importance [15,12], yet they correlate only weakly with adversarial vulnerability.

We show that these measures cannot deterministically identify edges whose removal maximally degrades ranking, motivating a learning-based estimation of edge impact.

Proposition 1 (Embedding Perturbation Formulation). *Let $e = (v_i, v_j) \in E$ be an edge in an undirected graph $G = (V, E)$. The marginal embedding perturbation on a target node v_t after removing e is $\Delta \mathbf{z}_{v_t}(e) = f_{GNN}(\mathbf{L}_{-e}, \mathbf{X}) - f_{GNN}(\mathbf{L}, \mathbf{X})$, where $\mathbf{L}_{-e}$ is the Laplacian after deleting e. For a linearized k-layer GNN such as GCN, $f_{GNN}(\mathbf{L}, \mathbf{X}) = \mathbf{A}^k\mathbf{X}\mathbf{W}$, yielding:*

$$\Delta \mathbf{z}_{v_t}(e) = \sum_{l=1}^{k} \left[(\mathbf{A}^{l-1})_{v_t,i} \mathbf{z}_{v_j}^{(k-l)} + (\mathbf{A}^{l-1})_{v_t,j} \mathbf{z}_{v_i}^{(k-l)} \right]$$

Hence, the embedding shift depends on multi-hop adjacency interactions $(\mathbf{A}^k)_{v_t,i}$ and $(\mathbf{A}^k)_{v_t,j}$ rather than first-order node statistics such as degrees or PageRank scores.

Theorem 1 (Non-Correlation of Structural Metrics and Rank Change). *For a connected graph $G = (V, E)$ and a fixed target node v_t, no monotonic function $g : \mathbb{R}^2 \to \mathbb{R}$ satisfies $g(d_i, d_j) \propto \Delta r(v_t, e)$ for all edges $e = (v_i, v_j) \in E$, where $\Delta r(v_t, e)$ denotes the rank change of v_t after removing e in a similarity-based retrieval model.*

Proof. The rank change $\Delta r(v_t, e)$ is determined by the derivative of the similarity $\mathrm{sim}(\mathbf{z}_{v_t}, \mathbf{q})$ with respect to the adjacency perturbation ΔA_{ij}:

$$\frac{\partial \mathrm{sim}(\mathbf{z}_{v_t}, \mathbf{q})}{\partial A_{ij}} \propto \frac{\partial \mathbf{z}_{v_t}}{\partial A_{ij}} = \sum_{k=1}^{K} (\mathbf{A}^{k-1})_{v_t,i} (\mathbf{X}\mathbf{W})_j + (\mathbf{A}^{k-1})_{v_t,j} (\mathbf{X}\mathbf{W})_i$$

This depends on non-linear multi-hops in $\mathbf{A}$, while degree/PageRank capture first-order connectivity. So, no monotonic mapping can generalize to all adjacency configurations.

2.2 Proposed Approach

The proposed framework formulates adversarial edge removal as a supervised estimation problem in which the goal is to approximate the mapping between local structural configurations and their induced ranking distortions. For each edge $e = (v_i, v_j)$, the model seeks to predict the magnitude of rank degradation that would result from its removal, denoted as $s(e) = \Delta r(v_t, e)$. This learning-based formulation replaces deterministic structural heuristics with a data-driven surrogate that infers sensitivity patterns directly from observed perturbation–response pairs. The central assumption is that the relationship between an edge's local embedding context and its contribution to retrieval stability is continuous but highly non-linear; thus, it can be learned through function approximation without requiring explicit gradient access or model retraining.

Formally, let $\mathcal{H}$ denote the space of edge representations that jointly encode topological and embedding-based properties, and let $\mathcal{S}$ represent the corresponding space of ranking degradation magnitudes. The learning objective is to approximate a function $f_\theta : \mathcal{H} \to \mathcal{S}$ such that $f_\theta(\mathbf{h}_e) \approx s(e)$ where f_θ is parameterized by θ and optimized to minimize the expected prediction error over the joint distribution of $(\mathbf{h}_e, s(e))$. This allows the model to generalize across unseen graphs by capturing invariant patterns of embedding sensitivity, providing an efficient surrogate for the intractable process of evaluating every possible edge perturbation. In our formulation, each edge $e = (v_i, v_j)$ is encoded as a local feature vector $\mathbf{h}_e$ that summarizes its contribution to the surrounding

embedding geometry. This vector captures both direction and magnitude in the latent space, reflecting the strength and orientation of the dependency between v_i and v_j. The removal of e perturbs message propagation paths and modifies higher-order structural correlations in $\mathbf{A}^k\mathbf{X}$; hence, $\mathbf{h}_e$ implicitly encodes multi-scale spectral interactions that govern the stability of embeddings under structural change.

To enhance expressiveness, $\mathbf{h}_e$ can be augmented with auxiliary structural descriptors such as node degrees: $\mathbf{h}_e = [\, AGG(\mathbf{z}_{v_i}, \mathbf{z}_{v_j}) \,\|\, d_{v_i}, d_{v_j} \,]$, where AGG denotes aggregation function, $\|$ is concatenation, d_{v_i} and d_{v_j} are node degrees. These additional statistics introduce global priors that complement the localized embedding information, allowing the model to reason jointly over geometric and structural information.

Edge Impact Estimator. Given a dataset $\mathcal{D} = \{(\mathbf{h}_e, s(e)) \mid e \in E_{\text{train}}\}$, the estimator learns a continuous approximation of the mapping from edge features to rank degradation. The expected risk minimization problem is defined as:

$$\min_{\theta} \; \mathbb{E}_{(\mathbf{h}_e, s(e)) \sim \mathcal{D}} \big[(f_\theta(\mathbf{h}_e) - s(e))^2 \big],$$

where f_θ is implemented as a Multi-Layer Perceptron (MLP) with nonlinear activation functions. The MLP structure enables the capture of high-order, non-additive dependencies between embedding and structural signals. From a functional perspective, f_θ approximates the sensitivity operator that maps local topological perturbations to embedding-space distortions, serving as an implicit estimator of $\partial \mathbf{z}_{v_t} / \partial A_{ij}$ without requiring explicit gradients from the retrieval model.

We use the listwise learning-to-rank formulation [2] as the training objective, encouraging the model to predict higher impact scores for edges that contribute more strongly to rank demotion. Since the normalized discounted cumulative gain (NDCG) is non-differentiable, we employ a differentiable pairwise logistic surrogate that approximates its behavior. For each target node v_t with incident edges $E_t = \{e_1, \ldots, e_n\}$, let $\mathcal{P}_{v_t} = \{(e_i, e_j) \mid s(e_i) > s(e_j)\}$ denote the ordered edge pairs based on ground-truth impact scores. The loss is defined as:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{v_t \in \mathcal{D}} \frac{1}{|\mathcal{P}_{v_t}|} \sum_{(e_i, e_j) \in \mathcal{P}_{v_t}} \log\big(1 + \exp(-(\hat{s}_{e_i} - \hat{s}_{e_j}))\big),$$

where $\hat{s}_{e_i} = f_\theta(\mathbf{h}_{e_i})$ represents the predicted edge score.

2.3 Attack Inference and Complexity

Once the estimator f_θ has been trained to approximate the mapping between edge representations and their induced ranking distortions, it can be used to perform targeted adversarial attacks at inference time. For a given target node v_t, the attacker evaluates all incident edges $\mathcal{E}_{v_t} = \{(v_t, u) \mid (v_t, u) \in E\}$ using the learned model to obtain predicted impact scores $\hat{s}(e) = f_\theta(\mathbf{h}_e)$. These scores quantify the estimated sensitivity of the retrieval ranking to the removal of each edge. The attack is then formulated as a discrete optimization problem under budget constraint. Specifically, the adversary seeks to select a subset of edges $\mathcal{E}^* \subseteq \mathcal{E}_{v_t}$ whose removal maximizes the cumulative impact:

$$\max_{\mathcal{E}^* \subseteq \mathcal{E}_{v_t},\, |\mathcal{E}^*| \le B} \sum_{e \in \mathcal{E}^*} \hat{s}(e)$$

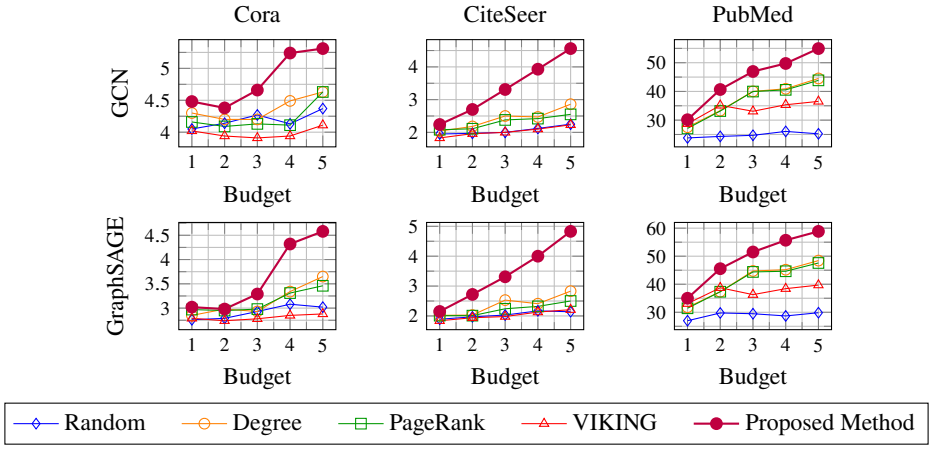

Fig. 1. Effect of edge-removal budget on demotion performance for different attack strategies.

This objective is efficiently approximated by ranking all incident edges in descending order of $\hat{s}(e)$ and selecting the top-k elements: $\mathcal{E}^* = \text{Top-}k\big(\{\hat{s}(e) \mid e \in \mathcal{E}_{v_t}\}\big), \quad k \leq B$ This inference mechanism constitutes a learned surrogate for the combinatorial search problem: $\max_{E' \subseteq E,\, |E'| \leq B} \sum_{e \in E'} s(e)$, which would otherwise require evaluating the true rank degradation $s(e)$ for every candidate edge. The proposed method reduces the attack complexity from $O(|E||V|)$ for exhaustive search to $O(|\mathcal{E}_{v_t}|)$ at inference time.

3 Experiments

Datasets. We evaluate our attack on three standard citation network benchmarks: `Cora` [8], `CiteSeer` [14], and `PubMed` [9]. Cora contains 2,708 nodes, 10,556 edges, and seven classes; CiteSeer has 3,327 nodes, 9,104 edges, and six classes; and PubMed includes 19,717 nodes, 88,648 edges, and three classes. Our code is available **here**.
Evaluation Setup. We train two GNNs, `GCN` [6] and `GraphSAGE` [5], using an unsupervised framework [12], to construct node representations on graphs. Cosine similarity is used as the ranking metric, and attack effectiveness is measured via `Average Rank Demotion (ARD)`, the mean drop in a target node's ranking after perturbation.
Baselines. We include several baselines: (1) *Degree*, which removes edges based on the degree values of neighbor nodes; (2) *PageRank* [7], which removes edges linked to nodes with the personalized PageRank score; (3) *Random*, which removes edges uniformly at random; (4) *VIKING* [4], a supervised gradient-based poisoning attack.
Findings. Our key observations are summarized as follows: **(1) Overall Performance:** The Average Rank Demotion (ARD) results are presented in Figure 1. Our proposed method consistently achieves higher rank demotion compared to all baselines. This trend holds not only in single-edge attack scenarios but also when the perturbation budget increases, demonstrating that the model effectively identifies high-impact edges even as the attack scope expands. Moreover, the performance gap between our method and the baselines widens with larger budgets, indicating that heuristic methods, despite their ability to target influential nodes, fail to select globally optimal edges for demotion. **(2) Attack Success Rate:** Table 1 reports the percentage of target nodes successfully demoted by each attack. While heuristic strategies such as *Degree* and *PageRank* perform

Table 1. Attack Success Rate (%) across datasets for GCN and GraphSAGE models.

Method	GCN			GraphSAGE		
	Cora	CiteSeer	PubMed	Cora	CiteSeer	PubMed
Degree	38%	61%	67%	36%	57%	70%
Page Rank	38%	58%	71%	36%	52%	73%
Viking	29%	47%	61%	29%	43%	65%
Proposed Method	**60%**	**81%**	**79%**	**53%**	**85%**	**81%**

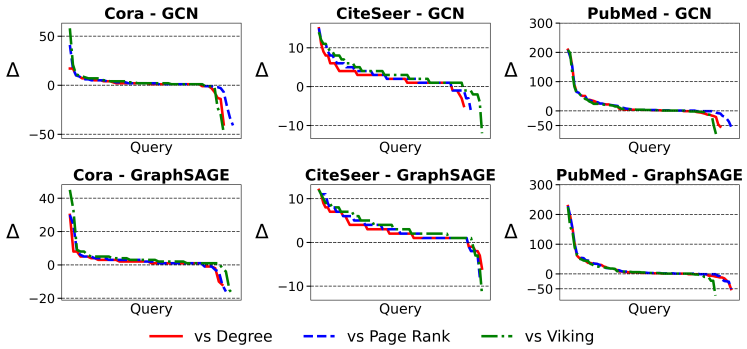

Fig. 2. Comparison of our proposed method vs. baselines in the Δ of helped and hurt queries.

comparably, the perturbation-based VIKING model underperforms across all datasets. Notably, success rates increase with graph size, suggesting that although heuristic methods are suboptimal, they become more effective in denser, highly connected graphs. Nevertheless, our proposed approach achieves the highest overall success rate across all settings, validating its robustness and scalability. **(3) Help–Hurt Analysis:** The help–hurt diagram in Figure 2 compares changes in target node ranks between our method and the top three baselines. Our model not only affects a larger proportion of nodes but also produces stronger demotion magnitudes. In particular, when the rank displacement (Δ) is analyzed, our method exhibits a substantially higher average demotion, confirming that it consistently identifies the most influential structural perturbations. Overall, these results demonstrate the efficiency of our method in identifying the most impactful edges, extending effectively to multi-edge attack scenarios, and maximizing attack performance in terms of demotion magnitude and success rate.

4 Concluding Remarks

We presented a learning-based framework for adversarial edge removal in graph-based retrieval systems. Our approach learns to estimate edge vulnerability directly from local embedding interactions, enabling black-box attacks that degrade retrieval rankings. Experiments on real-world datasets demonstrate superior performance over all baselines, achieving significant improvements in both rank demotion and attack success rate under perturbation budgets. These findings highlight the importance of data-driven modeling for exploiting adversarial sensitivity in graph structures. Future work will explore extending this framework with contrastive learning objectives for adaptive attack generation, as well as leveraging it for defense design in graph-based retrieval and recommendation.

References

1. Bryson, S., Davoudi, H., Golab, L., Kargar, M., Lytvyn, Y., Mierzejewski, P., Szlichta, J., Zihayat, M.: Robust keyword search in large attributed graphs. Inf. Retr. J. **23**(5), 502–524 (2020). https://doi.org/10.1007/S10791-020-09379-9, https://doi.org/10.1007/s10791-020-09379-9
2. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning. p. 129–136. ICML '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1273496.1273513, https://doi.org/10.1145/1273496.1273513
3. Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., Song, L.: Adversarial attack on graph structured data. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 1123–1132. PMLR (2018), http://proceedings.mlr.press/v80/dai18b.html
4. Gupta, V., Chakraborty, T.: VIKING: adversarial attack on network embeddings via supervised network poisoning. In: Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R.K., Reddy, P.K., Srivastava, J., Chakraborty, T. (eds.) Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part III. Lecture Notes in Computer Science, vol. 12714, pp. 103–115. Springer (2021). https://doi.org/10.1007/978-3-030-75768-7_9, https://doi.org/10.1007/978-3-030-75768-7_9
5. Hamilton, W.L., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1024–1034 (2017), https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html
6. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=SJU4ayYgl
7. Klicpera, J., Bojchevski, A., Günnemann, S.: Predict then propagate: Graph neural networks meet personalized pagerank. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=H1gL-2A9Ym
8. McCallum, A.K., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. Inf. Retr. **3**(2), 127–163 (2000). https://doi.org/10.1023/A:1009953814988, https://doi.org/10.1023/A:1009953814988
9. Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. AI Mag. **29**(3), 93–106 (2008). https://doi.org/10.1609/AIMAG.V29I3.2157, https://doi.org/10.1609/aimag.v29i3.2157
10. Sun, Y., Wang, S., Tang, X., Hsieh, T., Honavar, V.G.: Node injection attacks on graphs via reinforcement learning. CoRR **abs/1909.06543** (2019), http://arxiv.org/abs/1909.06543
11. Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., Jain, A.K.: Adversarial attacks and defenses in images, graphs and text: A review. Int. J. Autom. Comput. **17**(2), 151–178 (2020). https://doi.org/10.1007/S11633-019-1211-X, https://doi.org/10.1007/s11633-019-1211-x

12. Xu, Y., Huang, S., Zhang, H., Li, X.: Why does dropping edges usually outperform adding edges in graph contrastive learning? In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 21824–21832. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I20.35488, https://doi.org/10.1609/aaai.v39i20.35488
13. Yang, J., Yao, W., Zhang, W.: Keyword search on large graphs: A survey. Data Science and Engineering **6**(2), 142–162 (2021)
14. Yang, Z., Cohen, W.W., Salakhutdinov, R.: Revisiting semi-supervised learning with graph embeddings. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 40–48. JMLR.org (2016), http://proceedings.mlr.press/v48/yanga16.html
15. Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L.: Graph contrastive learning with adaptive augmentation. In: Leskovec, J., Grobelnik, M., Najork, M., Tang, J., Zia, L. (eds.) WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021. pp. 2069–2080. ACM / IW3C2 (2021). https://doi.org/10.1145/3442381.3449802, https://doi.org/10.1145/3442381.3449802
16. Zügner, D., Akbarnejad, A., Günnemann, S.: Adversarial attacks on neural networks for graph data. In: Guo, Y., Farooq, F. (eds.) Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018. pp. 2847–2856. ACM (2018). https://doi.org/10.1145/3219819.3220078, https://doi.org/10.1145/3219819.3220078
17. Zügner, D., Günnemann, S.: Adversarial attacks on graph neural networks via meta learning. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bylnx209YX