RSA-CR: Resisting Shilling Attacks in Citation Recommendation via Dumbbell Inductive Learning

Xiyue Gao[1], Yukai Liu[1], Zhuoqi Ma[1]*, Xiaotian Qiao[1], Hui Li[1], Cai Xu[1], Kunhua Zhang[1], Jiangtao Cui[1]

[1]School of Computer Science and Technology, Xidian University
Xi'an, Shaanxi, 710126, China

Abstract

Citation recommendation aims to provide researchers with the most relevant references for their manuscripts, helping them swiftly discover pertinent studies and bolster the reliability of their arguments. However, some individuals manipulate these recommendation systems by injecting false information, such as deliberately inflating the citation count of their own papers, to obtain favorable recommendations and ratings. This form of attack, commonly termed "shilling attack", is not only highly concealed but also has an unimaginable impact on all scientific research. To address this problem, we theoretically reveal the impact of shilling attacks on citation recommendation and propose three feasible resistance strategies: historical collaborations, significant citations and content constraints. Based on these insights, we introduce RSA-CR, a robust and hybrid citation recommendation algorithm resistant to shilling attacks. The algorithm constructs a two-layer academic graph and uses random and content generation strategies to initialize author and paper embeddings. Confidence-guided inductive aggregations based on collaboration and citation relationships are then performed at the author and paper sides, where author aggregation results directly influences the paper aggregation strength. Finally, recommendations are made by measuring the distances between the fused paper embeddings. The entire learning process resembles a dumbbell, hence termed "dumbbell inductive learning". Experiments on four academic datasets demonstrate that our method outperforms baselines in both effectiveness and robustness.

Code — https://github.com/kevinnevi/RSA-CR

Introduction

Academic papers are the most direct and effective way to understand technological progress (Sugiyama and Kan 2010). Currently, the number of academic papers is growing exponentially, with approximately 7.2 million papers published in 2022 alone, and this trend continues to rise annually (Curcic 2023). However, recent research indicates that regardless of the subject area, the groundbreaking trend of papers is decreasing (Park, Leahey, and Funk 2023). Under the dual impact of the continuous growth in the number of papers and the reduction of groundbreaking papers, citation recommendation faces greater challenges (Kozlov 2023).

Citation recommendation, particularly referring to global citation recommendation (Färber and Jatowt 2020), aims to provide researchers with the most pertinent references for their manuscripts. An excellent citation recommendation system can swiftly guide researchers to discover relevant studies, strengthen the reliability of argumentation, improve the quality and efficiency of research, and even promote the development of science (Kreutz and Schenkel 2022).

At present, citation recommendation systems have been widely used in academia. Existing methods fall into four categories (Ali et al. 2020): collaborative filtering (CF), content-based filtering (CBF), graph-based methods (GB), and hybrid methods. CFs recommend citations by capturing similarities between users or papers (Wu et al. 2022; He et al. 2017). CBFs achieve citation recommendation by extracting content information from papers (Son and Kim 2017). GBs construct homogeneous or heterogeneous graphs and explore auxiliary relationships for recommendation while alleviating data sparsity and cold start issues (Shahid et al. 2020; Guo et al. 2020). Hybrid methods combine multiple strategies to provide comprehensive recommendations (Gündoğan and Kaya 2022; Guo et al. 2022).

Although these methods achieve decent citation recommendations, most of them have a strong dependence on citation networks or ratings (Guo et al. 2020), making them vulnerable to attacks. Some manipulate these recommendation systems by injecting false information, like deliberately interfering with the citation count of their own papers, to obtain favorable recommendations. This form of attack is commonly known as "shilling attack" (Su, Zeng, and Chen 2005; Shahid et al. 2020). In addition, existing research indicates that, due to conservative tendencies and herd mentality among researchers, papers frequently cited in one year tend to maintain high citation rates in the subsequent year (Chu and Evans 2021; Kurdi 2021). This phenomenon, referred to as "class solidification" in the scientific community, further exacerbates the impact of shilling attacks.

Current robust recommendation algorithms against shilling attacks mostly focus on item recommendation, employing strategies such as suspicious user detection, introducing trust models, and multi-dimensional trust evaluation (Gunes et al. 2014; Si and Li 2020; Ye et al. 2023). How-

*Corresponding author.

ever, research on robust recommendation against shilling attacks in citation recommendation is limited, likely due to the unique characteristics of shilling attacks in this domain:

- ***Data complexity.*** Citation recommendation involves complex data, including citation relationships, paper content, author information, etc., and the number of papers is growing rapidly every year. Therefore, algorithms should extract useful information from these complex data to resist shilling attacks and promptly recommend newly published papers in a high-throughput environment (Hamilton, Ying, and Leskovec 2017).
- ***Attack concealment.*** Item recommendation systems monitor user behavior, but the complexity and professionalism of the academic field make shilling attacks more covert and harder to distinguish from regular activities, increasing the difficulty of detection and defense.
- ***Significant impact.*** Shilling attacks in item recommendation affect market fairness and consumer choice, while in citation recommendation, they distort academic research perceptions, influence evaluation standards, funding, and even the direction of scientific development.

To solve the above problems, we first theoretically reveal the impact of shilling attacks on citation recommendation and propose three feasible resistance strategies: historical collaborations, significant citations, and content constraints. Based on these insights, we propose RSA-CR, a robust and hybrid citation recommendation algorithm resistant to shilling attacks. The algorithm first constructs a dual-layer academic graph, using random and content generation strategies to initialize embeddings for authors and papers, respectively. Then, sampling and confidence-guided inductive aggregations based on collaboration and citation relationships are performed separately for authors and papers, where author aggregation results directly influences the paper aggregation strength. Inductive and confidence-based aggregations can be used to quickly extract features of newly published papers and enhance robustness. Finally, recommendations are made by measuring the distances between the fused paper embeddings. The entire learning process resembles a dumbbell, hence referred to as dumbbell inductive learning (DIL). To validate the effectiveness of RSA-CR, we conducted comparative experiments on four academic datasets. The experimental results show that our method outperforms existing approaches in both recommendation effectiveness and robustness. Our contributions can be summarized as follows:

- We propose a novel problem: shilling attacks in citation recommendation. Though existent in academia, to our knowledge, there has been no systematic research on it.
- We theoretically analyze shilling attack characteristics in academia and propose three strategies.
- We design a robust and hybrid citation recommendation algorithm, RSA-CR, which employs inductive and confidence-guided aggregations to promptly extract features of new papers and enhance robustness.
- We create a Shilling Attack Academic Dataset (Saaca), with evaluations on it and three other datasets demonstrating RSA-CR's superior effectiveness and robustness.

Theoretical Guarantees

Preliminaries

Problem 1 *Given a candidate manuscript p_c with few or no citations, the goal of **citation recommendation** task (global (Färber and Jatowt 2020)) is to devise a prediction function f that estimates the likelihood of any paper p_r in the scientific paper database $D = \{p^i \mid i = 1, 2, \ldots, |D|\}$ being cited by p_c:*

$$P_D^k \leftarrow Top_k\{f(p_r \in D \mid p_c \notin D)\} \,. \tag{1}$$

Here, $|D|$ represents the total number of papers in D, and P_D^k signifies the subset of the top k papers within D most likely to be cited.

Assumption 1 *The prediction function f is assumed to be an unbiased estimator of the true citation probability distribution η. This means that for all pairs (p_1, p_2) in D where p_1 cites p_2, the expected value $\mathbb{E}[f(p_2 \mid p_1)] = \eta(p_2 \mid p_1)$. Additionally, f is consistent: as the size of the database D increases, the estimates $f(p_2 \mid p_1)$ converge in probability toward the true citation probabilities $\eta(p_2 \mid p_1)$. Specifically, for all $\psi > 0$,*

$$\lim_{|D|\to\infty} \mathbb{P}\left(|f(p_2 \mid p_1) - \eta(p_2 \mid p_1)| < \psi\right) = 1 \,. \tag{2}$$

Generally, the factors influencing η in a paper encompass details such as the authors, keywords, content, citations, and locations. Prior studies have established that citation relationships elucidate the similarities and connections among papers, which helps improve the accuracy and speed of recommendation systems (West, Wesley-Smith, and Bergstrom 2016; Liu et al. 2022).

Assumption 2 *Citation relationships can form a network $G = (V, E)$ derived from D, where V includes all papers in D, and E consists of the citation links among these papers. We hypothesize that G impacts η directly or indirectly. In other words, we assume that the prediction process is inevitably affected by these citation relations. Although ignoring them may enhance resistance to shilling attacks, it will significantly reduce the recommendation effectiveness.*

Shilling Attacks

Currently, certain individuals manipulate recommendation systems via shilling attacks (injecting false information). To boost robustness, it is crucial to clarify the intent, type, and implementation of such attacks. Nearly all shilling attacks aim to push or nuke the popularity of specific targets to gain benefits over competitors. Some attacks aim to increase the popularity of certain targets (push attacks), while others aim to decrease it (nuke attacks) (Mobasher et al. 2007). Shilling attacks in citation recommendations typically aim to boost the popularity of specific papers and are thus considered push attacks, defined as follows:

Definition 1 *In citation recommendation, a **shilling attack** refers to introducing a set of attacking papers $P_a = \{p_a^i \mid i = 1, 2, \ldots, |P_a|\}$ into the scientific paper database D. Each p_a^i has a citation list $C(p_a^i)$ that includes target citations P_{ta} and authentic citations P_{au}^i. P_{ta} is largely consistent across all papers in this attacking set, while P_{au}^i is unique to each p_a^i and disjoint from P_{ta}.*

Assumption 1 and 2 underlie our study of shilling attacks' impact on citation recommendations. To evaluate this quantitatively, we introduce the following additional assumption:

Assumption 3 *Assume all papers in D are embedded within a high-dimensional Euclidean space S_G. In this space, the Euclidean distance between any two papers is directly proportional to the likelihood of one citing the other: the closer the two papers, the higher the citation probability. Specifically, if there is a candidate manuscript p_c and a random paper p_r, the citation probability from p_c to p_r is*

$$\eta_{S_G}(p_r \mid p_c) = e^{-\gamma d_{S_G}(p_r, p_c)} , \tag{3}$$

where γ is a scaling factor representing sensitivity to distance, and $d_{S_G}(p_r, p_c)$ denotes the Euclidean distance between p_r and p_c in space S_G.

Previous research described citation networks as scale-free, with node degrees following a power-law distribution. However, recent work (Brzezinski 2014) suggests the degree distribution more closely follows a log-normal distribution. We support this finding; our analysis on four academic datasets indicates that both node degrees and shortest path distances in citation networks are more consistent with a log-normal distribution. Given this observation, we propose the following assumption:

Assumption 4 *Node distances in S_G follow a log-normal distribution. Thus, the expected distances from p_c to p_r is*

$$E\left[d_{S_G}(p_r, p_c)\right] = e^{\mu_{S_G} + \delta^2_{S_G}/2} . \tag{4}$$

After a shilling attack P_a, the space transforms from S_G to $S_{\tilde{G}} = G \cup P_a$. We assume P_a is introduced by the authors of p_r and that $\forall p_a \in P_a$, $d(p_r, p_c) = \epsilon$, where ϵ is a small value independent of $d_{S_{\tilde{G}}}(p_a, p_c)$. We further assume there is a citation neighborhood $S(p_c)$ around p_c, where all nodes are potential citation targets for p_c. The probability that an attacking paper p_a falls within $S(p_c)$ is denoted as p. In $|P_a|$ independent Bernoulli trials, the number of papers falling within $S(p_c)$ follows a binomial distribution. Thus, the probability that at least one paper falls within $S(p_c)$ is $1-(1-p)^{|P_a|}$, while the probability that no attacking papers fall within $S(p_c)$ is $(1-p)^{|P_a|}$. Under these assumptions, since these attacking nodes typically cite core nodes more actively, we examine the expected post-attack distance $d_G(p_r, p_c)$ when $d_{S_{\tilde{G}}}(p_a, p_c) \leq d_{S_G}(p_r, p_c) - \epsilon$:

$$\begin{aligned} E\left[d_{S_{\tilde{G}}}(p_r, p_c)\right] &= E\left[\min(d_{S_{\tilde{G}}}(p_a, p_c) + \epsilon,\ d_{S_G}(p_r, p_c))\right] \\ &\approx E\left[d_{S_{\tilde{G}}}(p_a, p_c) + \epsilon\right] \\ &= \left(1-(1-p)^{|P_a|}\right) K + (1-p)^{|P_a|} L + \epsilon \\ &\approx \left(1-(1-p)^{|P_a|}\right) K \\ &\quad + (1-p)^{|P_a|} E[d_{S_G}(p_r, p_c)] + \epsilon , \end{aligned} \tag{5}$$

where K and L denote the expected distances within and outside the $S(p_c)$ region, respectively.

To quantify the impact of the shilling attack, we compare the expected citation probabilities before and after the attack. We define $A = e^{\mu_{S_G} + \delta^2_{S_G}/2}$ as a fixed value directly related to the characteristics of the citation network. This difference is given by:

$$\begin{aligned} \Delta E\left[P_a\right] &= \mathbb{E}\left[\frac{\eta_{S_{\tilde{G}}}(p_r \mid p_c) - \eta_{S_G}(p_r \mid p_c)}{\eta_{S_G}(p_r \mid p_c)}\right] \\ &= e^{\gamma A}\left(e^{-\gamma\left((1-(1-p)^{|P_a|})K + (1-p)^{|P_a|}A + \epsilon\right)} - 1\right) \\ &= e^{-\gamma\left((1-(1-p)^{|P_a|})K + (1-p)^{|P_a|}A + \epsilon - A\right)} - 1 \\ &= e^{\gamma\left((1-(1-p)^{|P_a|})(A-K) - \epsilon\right)} - 1 . \end{aligned} \tag{6}$$

From this, we can derive the following proposition:

Proposition 1 *Given Assumptions 1, 2, 3, and 4, the change in citation probability after a shilling attack is influenced by several factors. Key factors include the scaling factor, the number of attacking papers, and their distribution. The scaling factor γ is generally directly related to the distance between nodes in space S_G. $|P_a|$ and p are parameters directly related to the attacking papers, determining the strength and accuracy of the attack. As $|P_a|$ and p increase, the change in probability also becomes greater. The parameter ϵ indicates the proximity of the attacking papers to the target paper. The smaller the ϵ, the more pronounced the attack effect.*

Based on this proposition, we substitute specific numerical parameters for numerical verification. Assuming the node distances in S_G correspond to the shortest paths in the original network, γ can generally be set to around 0.1. Based on an analysis of μ and θ in four real-world citation networks, we estimate that falls within the range of 5 to 15. For our calculations, we assume $A = 10$ and $K = 3$. p plays a crucial role, especially when attacking papers actively cite high-degree nodes, causing p to be high. When $p = 0.1$ and $|P_a| = 50$, according to Eq. 6, the recommendation probability may increase by 80%; when $p = 0.01$ and $|P_a| = 100$, it may increase by 40%. This confirms that shilling attacks can significantly impact citation recommendations.

Resisting Shilling Attacks

Problem 2 *The goal of **resisting shilling attacks** is to ensure that the citation recommendation system, while meeting the basic requirements (Problem 1), experiences minimal changes in recommendation results when subjected to a series of shilling attacks:*

$$P_D^k \sim P^k_{D \cup \bigcup_{i=1}^n P_a^i} , \tag{7}$$

where $D \cup \bigcup_{i=1}^n P_a^i$ represents the scientific paper database after n shilling attacks, and $\sim$ indicates that the two recommendation results are similar.

Proposition 1 suggests two solutions to Problem 2: First, improve data quality by reducing attacking papers $|P_a|$. Although attack detection techniques exist in item recommendation, they are rare in citation recommendation and often inapplicable due to differing attack types. Second, redefine the distance metric in space $S_{\tilde{G}}$ to improve robustness, which is the focus of this study. Typically, the simplest distance definition in space $S_{\tilde{G}}$ is the shortest path in the original network. However, it treats all relationships equally, resulting in suboptimal resistance to shilling attacks. As shown

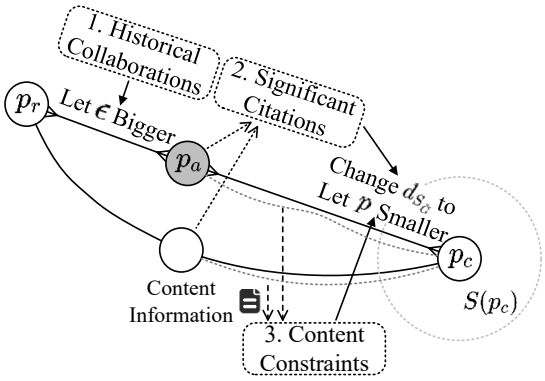

Figure 1: Three strategies for resisting shilling attacks.

in Figure 1, we propose integrating three types of additional information to refine this metric in $S_{\tilde{G}}$. This will increase the path length through attacking papers P_a while minimally affecting normal papers, thus enhancing overall resistance.

- ***Historical collaborations.*** Generally, the authors involved in a shilling attack are familiar with the authors of the target paper. By measuring the degree of familiarity between these authors through historical collaborations or other additional information, the distance between P_a and P_r in $S_{\tilde{G}}$ can be increased, thereby increasing the ϵ value. In comparison, normal citations are less affected, thus raising the system's resistance to shilling attacks.
- ***Significant Citations.*** A major issue with current methods is their inability to distinguish the impact of different citations. Therefore, to resist shilling attacks, the distance function $d_{S_{\tilde{G}}}$ can be modified using information from the cited papers. Additional paper features along the citation path, such as journal ratings and author reputation, can be used to bring P_a closer to high-credibility papers, thereby indirectly reducing p and enhancing resistance.
- ***Content constraints.*** A notable characteristic of shilling attack papers is their ability to significantly alter network connectivity by linking distant nodes. Therefore, the third strategy is to utilize additional contextual information, such as paper content, to help identify unreasonable cross-domain connections. Widening the content gap between P_a and unrelated papers, we can also reduce the parameter p and boost resistance to shilling attacks.

RSA-CR Algorithm

Overall Architecture

RSA-CR is a robust and hybrid citation recommendation algorithm. Its robustness lies in its ability to resist shilling attacks, while its hybrid nature comes from leveraging various information sources, including graphs, content, and context, for recommendations. The overall process is as follows: 1) *Middleware Extraction Phase*: Extract intermediate data items from the original scientific paper database D, including initial representations of authors and papers, dual-layer academic graphs, and additional author and location context information. 2) *DIL Feature Fusion Phase*: Perform confidence-guided inductive aggregation based on collaboration and citation relationships on both sides of the dual-layer academic graph. 3) *Citation Recommendation Phase*: Generate citation recommendations through distance measurement, using negative sampling (Mikolov et al. 2013) and gradient descent to learn model parameters. After training, RSA-CR can provide robust global citation recommendations for manuscripts with few or no citations.

Middleware Extraction

A paper consists of several components, including the title, authors, abstract, keywords, body, and references. Author information provides details on ownership and collaboration, while the title, abstract, and body offer insights into the paper's content. This data is crucial for citation recommendation and resistance to shill attacks. Therefore, the first step is to extract intermediate data items from the original paper, which we call "middleware".

Initial embeddings of authors and papers Traditional graph embedding-based recommendation algorithms are typically transductive, requiring all nodes to be present during training. However, academic paper networks are highly dynamic, with new papers appearing daily. Existing methods struggle with unobserved nodes during training, necessitating recommendation algorithms with inductive representation capabilities. This involves learning stable aggregation functions during training to provide recommendations for unseen nodes (Hamilton, Ying, and Leskovec 2017), slightly slowing down training speed but offering notable benefits.

For subsequent inductive learning, input embeddings are crucial. RSA-CR initializes author embeddings randomly and paper embeddings with doc2vec (Le and Mikolov 2014) to capture semantic information from titles, abstracts, keywords, and body. These content-based unsupervised embeddings enable the recommendation algorithm to fully incorporate semantic information, enhancing recommendation quality and resilience against shilling attacks.

Dual-layer academic graph Then, we extract relationships (collaborations, citations, authorships) from the collection of papers and construct a dual-layer academic graph. This graph, denoted $G_{da} = (V, E, T)$, is heterogeneous, where each node $v \in V$ and each link $e \in E$ are associated with mapping functions $\phi(v) : V \to T_V$ and $\phi(e) : E \to T_E$. Here, $T_V = \{\text{author}, \text{paper}\}$ and $T_E = \{\text{collaboration}, \text{authorship}, \text{citation}\}$ denote sets of node and link types. Each link type t_e is bound to a mapping function specifying the node types of its endpoints $\varphi(t_e) : T_E \to T_V$, where $\varphi(\text{collaboration}) = [\text{author}, \text{author}]$, $\varphi(\text{authorship}) = [\text{author}, \text{paper}]$, and $\varphi(\text{citation}) = [\text{paper}, \text{paper}]$.

Additional contextual information Context can provide a more comprehensive evaluation of papers, aiding selection of high-credibility papers and identification of significant citations. This includes optional details on authors (academic background, field of expertise, reputation, withdrawal record), papers (journal/conference), and institutions (research focus and academic influence), used to compute confidence scores for significant citations.

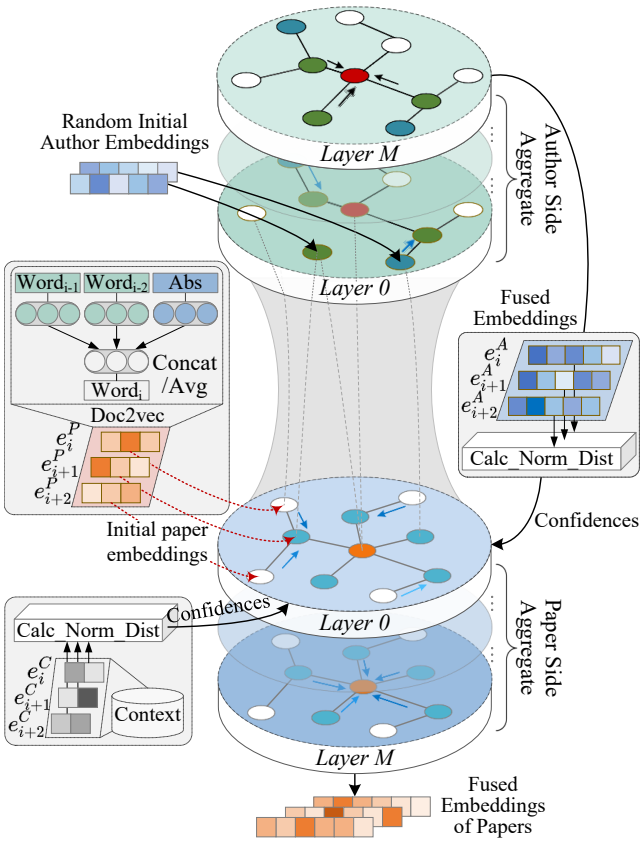

Figure 2: Illustration of dumbbell inductive learning.

Dumbbell Inductive Learning

Figure 2 shows the DIL structure, performing feature fusion over a dual-layer academic graph (dashed lines for authorship). It starts from the upper green (author) layers, proceeds to the lower blue (paper) layers, and ultimately outputs fused, robust paper embeddings resistant to shilling attacks.

Author side learning The author-side graph is essentially a weighted graph, where nodes represent authors, links denote collaboration relationships, and link weights reflect the number of past collaborations. Each node is initialized with a random embedding. The training objective of DIL is to learn two sets of M aggregation functions, one for the author side and one for the paper side, with each function corresponding to a specific layer in DIL.

During forward propagation, to obtain the fused representations of target nodes, the neighborhood set of these nodes must first be randomly sampled. The maximum depth of the neighborhood set is M, with each layer of DIL corresponding to a specific depth. For the l-th layer on the author side, the feature fusion process involves moving from the neighborhood set at $M-l$ hops to the set at $M-l-1$ hops. For a target author node v, the fusion process at the l-th layer is:

$$h_{a,v}^{(l+1)} = \mathrm{LN}\big(W_l^a \cdot \mathrm{CAT}\big(\mathrm{MEAN}\big(h_{a,u}^{(l)}, \forall u \in \mathcal{N}_{co}(v)\big), h_{a,v}^{(l)}\big)\big), \tag{8}$$

where $h_{a,v}^{(l)}$ is the input embedding of node v at layer l, $\mathcal{N}_{co}(v)$ denotes the set of neighbors connected via collaboration links, MEAN is the element-wise mean of the vectors for aggregating neighbor information, CAT is the concatenation operation, W_l^a is the weight matrix, and LN stands for layer normalization (Ba, Kiros, and Hinton 2016). LN not only provides the non-linear mapping capabilities of activation functions but also helps mitigate gradient explosion issues and enhances training stability by normalizing the distribution of each sample across each feature dimension.

Paper side learning Learning on the paper side builds on author side learning. First, we calculate the inter-paper confidences at the paper level using the fused embeddings from the author side and contextual information. Specifically, we define two types of confidence between paper nodes u and v: $\mathrm{CONF}_{hc}(u,v)$ and $\mathrm{CONF}_{sc}(u,v)$. $\mathrm{CONF}_{hc}(u,v)$, derived from historical cooperation, is calculated as follows:

$$\mathrm{CONF}_{hc}(u,v) = \sigma\Big(\mathrm{LSTM}\left(h_{a,i}^M, \forall i \in \mathcal{N}_{wr}(u)\right) \cdot \mathrm{LSTM}\left(h_{a,j}^M, \forall j \in \mathcal{N}_{wr}(v)\right)^T / \sqrt{d_k}\Big), \tag{9}$$

where $h_{a,i}^M$ signifies the output representation of node i on the author side, $\mathcal{N}_{wr}(u)$ denotes the ordered neighbors (aligned with the author order) associated with the authorship link type of node u. The use of LSTM as an aggregator ensures the retention of author order during feature fusion. d_k represents the dimension of fused author embeddings z_a, and σ denotes the sigmoid function. $\mathrm{CONF}_{sc}(u,v)$ is computed based on the significant citations strategy mentioned above, leveraging additional contextual information:

$$\mathrm{CONF}_{sc}(u,v) = \sigma\Big(\mathrm{MLP}\left(c(u)\right) \cdot \mathrm{MLP}\left(c(v)\right)^T / \sqrt{d_c}\Big), \tag{10}$$

where $c(u)$ and $c(v)$ are contextual information vectors for paper nodes u and v, MLP denotes a multi-layer perceptron used to process this contextual information with dimension d_c, and σ is the sigmoid function.

Paper-side feature fusion is similar to that at the author level, but the input embeddings are generated using doc2vec for content constraint. Additionally, the fusion process requires weighting based on the confidences. The specific fusion process for a paper node v at the l-th layer is as follows:

$$h_{p,v}^{(l+1)} = \mathrm{LN}\left(W_l^p \cdot \mathrm{CAT}\left(\mathrm{MEAN}\left(\mathrm{CONF}(u,v) \cdot h_{p,u}^{(l)}, \forall u \in \mathcal{N}_{ci}(v)\right), h_{p,v}^{(l)}\right)\right), \tag{11}$$

where $\mathrm{CONF}(u,v)$ denotes the weighted average of historical collaboration confidence and significant citation confidence, with α used as the weighting parameter:

$$\mathrm{CONF}(u,v) = \alpha \cdot \mathrm{CONF}_{sc}(u,v) + (1-\alpha) \cdot \mathrm{CONF}_{hc}(u,v). \tag{12}$$

When propagation reaches maximum depth M, we obtain the final fused embedding $h_{p,v}^M$ for paper v.

Recommendation and Optimization

The DIL training process starts with learning the parameters on the author side, then on the paper side. We use negative sampling and Bayesian Personalized Ranking (BPR)

Dataset	Model	HR@5	HR@10	HR@20	NDCG@5	NDCG@10	NDCG@20
ACL-OCL	LINE	0.1061	0.1275	0.1546	0.0870	0.0940	0.1008
	Paper2vec	0.1571	0.2086	0.2911	0.1197	0.1361	0.1568
	Hyperdoc2vec	0.1411	0.2164	0.3229	0.0950	0.1192	0.1459
	Graph2Gauss	0.1489	0.2532	0.4111	0.0947	0.1285	0.1681
	JGCF	0.5345	0.5981	0.6679	0.4398	0.4608	0.4798
	SGFCF	0.4971	0.5922	0.6629	0.4971	0.5356	0.5602
	RSA-CR	**0.7188**	**0.8371**	**0.9222**	**0.5612**	**0.5998**	**0.6214**
unarXive	LINE	0.5900	0.6000	0.6083	0.5691	0.5726	0.5749
	Paper2vec	0.6400	0.7166	0.8100	0.5522	0.5771	0.6009
	Hyperdoc2vec	0.1233	0.2150	0.3400	0.0715	0.1008	0.1320
	Graph2Gauss	0.3050	0.4250	0.5750	0.2146	0.2527	0.2901
	JGCF	0.2561	0.2886	0.3638	0.2277	0.2380	0.3073
	SGFCF	0.2740	0.3129	0.3701	0.1428	0.1519	0.1648
	RSA-CR	**0.7600**	**0.8100**	**0.8450**	**0.7258**	**0.7421**	**0.7506**
Aminer	LINE	0.5513	0.5933	0.6525	0.5001	0.5135	0.5283
	Paper2vec	0.5979	0.7190	0.8232	0.4699	0.5090	0.5354
	Hyperdoc2vec	0.1021	0.1849	0.3186	0.0610	0.0875	0.1210
	Graph2Gauss	0.3029	0.4192	0.5706	0.2202	0.2577	0.2958
	JGCF	0.6496	0.6808	0.7202	0.5882	0.5985	0.6084
	SGFCF	0.1435	0.2489	0.6009	0.0806	0.1162	0.1761
	RSA-CR	**0.8556**	**0.9053**	**0.9440**	**0.7427**	**0.7474**	**0.7689**
Saaca	LINE	0.2973	0.3324	0.3703	0.2345	0.2459	0.2555
	Paper2vec	0.0498	0.0637	0.0969	0.0402	0.0447	0.0529
	Hyperdoc2vec	0.1067	0.1903	0.3030	0.0641	0.0909	0.1193
	Graph2Gauss	0.0396	0.0667	0.1143	0.0254	0.0341	0.0459
	JGCF	0.3096	0.4297	0.5460	0.1768	0.2135	0.2467
	SGFCF	0.4769	0.5820	0.6933	0.4347	0.4843	0.5294
	RSA-CR	**0.7081**	**0.8728**	**0.9464**	**0.4740**	**0.5279**	**0.5468**

Table 1: Recommendation performance on different datasets.

loss (Rendle et al. 2009) for both sides to improve the generalization ability of the model. Positive instances are node pairs connected by collaboration (authors) and citation (papers) relationships, while negative instances are randomly selected unconnected pairs. The BPR loss penalizes the model if the predicted similarity for positive instances is lower than that for negative instances:

$$\mathrm{L}_{bpr} = -\sum_{(u,v,w)\in D} \ln \sigma\left(\hat{x}_{uv} - \hat{x}_{uw}\right) , \qquad (13)$$

where $(u, v, w) \in D$ is a triplet in training set D, with u as target, v as positive, and w as negative sample. $\hat{x}_{uv}$ and $\hat{x}_{uw}$ represent the predicted scores for the positive and negative pairs, respectively. The predicted score is calculated as the inner product of the node embeddings: $\hat{x}_{uv} = h^M_{a/p,u} \cdot h^M_{a/p,v}$. Training uses Adam optimizer with a 0.001 learning rate. After training, given a candidate manuscript, RSA-CR can compute the citation score for any paper. The top k papers with the highest scores are the final recommendations.

Experiment

Datasets and Experiment Setup

We conduct experiments on four datasets: ACL-OCL (Rohatgi et al. 2023), unarXive (Saier and Färber 2020), AMiner (Tang et al. 2008), and Saaca (containing papers retracted due to fraudulent peer review, served as a proxy for shilling attack detection). All test papers are published after training papers to avoid hindsight bias. We compare RSA-CR with six baselines: LINE (Tang et al. 2015), Paper2vec (Ganguly and Pudi 2017), Hyperdoc2Vec (Han et al. 2018), Graph2Gauss (Bojchevski and Günnemann 2018), JGCF (Guo et al. 2023) and SGFCF (Peng et al. 2024). Recommendation quality is assessed with NDCG@K and HR@K, while robustness is measured by Attack Hit Ratio (AHR), indicating manipulated citations in top ranks, and Prediction Shift (PS), which quantifies the average absolute change in predictions before and after the attack, as defined by:

$$\mathrm{PS} = \frac{1}{N}\sum_{i=1}^{N} \left|\mathrm{Score}_i^{\text{after}} - \mathrm{Score}_i^{\text{before}}\right| , \qquad (14)$$

where N is the total number of citations, and the scores refer to the predicted citation score of paper i pre- and post-attack. Both PS and AHR are inversely proportional to the robustness of the algorithm against such attacks, with lower values indicating enhanced resistance to manipulation.

Recommendation Performance Analysis

We first evaluate RSA-CR on standard citation recommendation without shilling attacks. Table 1 shows the comparison between RSA-CR and all baselines. (i) RSA-CR consistently outperforms all baselines across four datasets, likely due to its effective integration of multimodal information, including textual content, network structure, and contextual data, combined through inductive learning. (ii) Among the

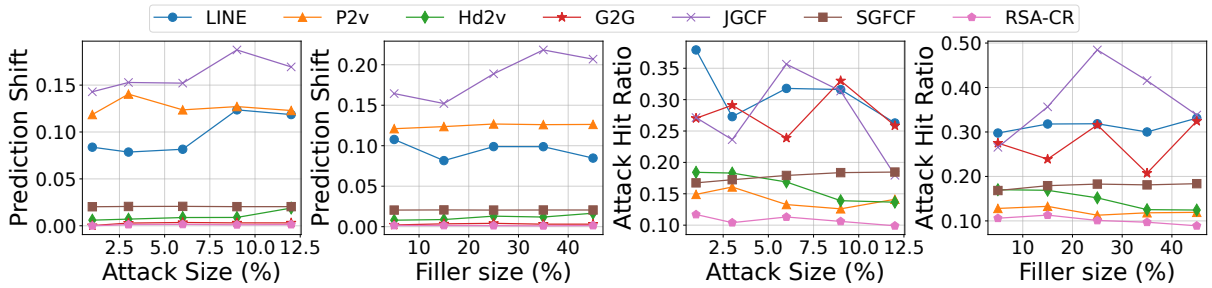

Figure 3: Prediction Shift and Hit Ratio for varying attack/filler sizes.

baselines, Paper2vec and JGCF perform best. Paper2vec integrates textual and citation information through distributed citation context representations, while JGCF employs Jacobi polynomial bases with frequency decomposition. Both are hybrid recommenders that leverage their strengths, demonstrating better adaptability across datasets. (iii) All methods perform worst on the Saaca dataset. However, RSA-CR shows the most significant improvement over baselines, likely due to the presence of retracted papers with low citation associations. RSA-CR consistently achieves high accuracy, highlighting its adaptability to dataset challenges.

Robustness Analysis

We conduct robustness experiments to evaluate resistance to shilling attacks. Figure 3 compares AHR/PS of RSA-CR and six baselines on AMiner with fixed 6% attack size (injected attacking papers) and varying filler sizes (non-target citations per attacking paper), and constant 15% filler size with varying attack sizes. RSA-CR outperforms baselines, maintaining stable AHR/PS across settings. This demonstrates RSA-CR's ability to filter out attacked items while maintaining normal recommendations, offering better robustness against shilling attacks. In contrast, most algorithms show higher PS values than RSA-CR across settings, indicating vulnerability, while AHR fluctuations result from large-scale attacks creating a clustering effect that reduces impact.

Ablation Study

Variants	HR@10	NDCG@10	AHR@10
RSA-CR	**0.8728**	**0.5279**	**0.0789**
w/o CONF$_{hc}$	0.7661	0.4436	0.0881
w/o CONF$_{sc}$	0.6961	0.3921	0.0968
w/o CONF$_{cc}$	0.7878	0.4363	0.0874

Table 2: Ablation Study on Saaca Dataset.

We conduct ablation experiments on the three resistance strategies in RSA-CR to assess the effectiveness of each component. The tested variants are: 1) without historical collaboration confidence (w/o CONF$_{hc}$), 2) without significant citations (w/o CONF$_{sc}$) and 3) without content constraints (w/o CONF$_{cc}$), while relying on other components optimal weighting. Table 2 presents the performance of RSA-CR and its variants on the Saaca dataset. Results indicate that all components are essential—removing any of them results in decreased performance to varying extents. We found that the variant lacking significant citation confidence exhibits the most substantial performance drop, highlighting the critical importance of high-quality citations. Significant papers written by high-reputation authors often act as network hubs, bringing related nodes closer and effectively mitigating the impact of low-quality citations and shilling attacks.

Case Study

Model	AHR@1	AHR@5	AHR@10
LINE	0.0331	0.0651	0.1375
Paper2vec	0.3668	0.7668	0.8319
Hyperdoc2vec	0.0340	0.0632	0.1032
Graph2Gauss	0.0239	0.0459	0.0839
JGCF	0.1751	0.4098	0.6442
SGFCF	0.0205	0.1250	0.2883
RSA-CR	**0.0044**	**0.0163**	**0.0789**

Table 3: Case Study on Saaca dataset.

The Saaca dataset features hypothetical labels for shilling attack targets. Based on this, we conduct a case study to further evaluate RSA-CR's resilience to such attacks. As shown in Table 3, although these target papers had high citation counts prior to retraction, RSA-CR consistently achieves the lowest AHR values, demonstrating strong robustness against shilling attacks. In contrast, baselines such as Paper2vec and JGCF show significantly higher AHR values. While both models perform well in terms of recommendation accuracy, they are highly sensitive to shilling attacks. This case highlights the prevalence of shilling attacks in academic domain and the vulnerability of baseline algorithms, while proving the effectiveness and robustness of RSA-CR.

Conclusion

We tackle shilling attacks in citation recommendation through theoretical analysis and propose RSA-CR, a robust algorithm that leverages a dual-layer academic graph and confidence-guided aggregation to resist such attacks effectively. Experiments on four datasets demonstrate that RSA-CR outperforms existing algorithms by approximately 24% in accuracy and exhibits a robustness improvement of about 15% against shilling attacks compared to baselines.

Acknowledgments

This work was supported by the Jing-Jin-Ji Regional Integrated Environmental Improvement-National Science and Technology Major Project of Ministry of Ecology and Environment of China (No. 2025ZD1200600), the National Natural Science Foundation of China under Grant Nos. 62302370, 62202360, 62302356, 62472340, and 62372352, and the CCF-ALIMAMA TECH Kangaroo Fund (No. CCF-ALIMAMA OF 2025007).

References

Ali, Z.; Kefalas, P.; Muhammad, K.; Ali, B.; and Imran, M. 2020. Deep learning in citation recommendation models survey. *Expert Systems with Applications*, 162: 113790.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bojchevski, A.; and Günnemann, S. 2018. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. In *International Conference on Learning Representations*.

Brzezinski, M. 2014. Power Laws in Citation Distributions: Evidence from Scopus. *SSRN Electronic Journal*.

Chu, J. S.; and Evans, J. A. 2021. Slowed canonical progress in large fields of science. *AAAI conference on artificial intelligence*, 118(41): e2021636118.

Curcic, D. 2023. Number of Academic Papers Published Per Year. https://wordsrated.com/number-of-academic-papers-published-per-year/.

Färber, M.; and Jatowt, A. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4): 375–405.

Ganguly, S.; and Pudi, V. 2017. Paper2vec: Combining graph and text information for scientific paper representation. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*, 383–395. Springer.

Gündoğan, E.; and Kaya, M. 2022. A novel hybrid paper recommendation system using deep learning. *Scientometrics*, 127(7): 3837–3855.

Gunes, I.; Kaleli, C.; Bilge, A.; and Polat, H. 2014. Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42: 767–799.

Guo, H.; Shen, Z.; Zeng, J.; and Hong, N. 2022. Hybrid Methods of Bibliographic Coupling and Text Similarity Measurement for Biomedical Paper Recommendation. *Studies in Health Technology and Informatics*, 290: 287–91.

Guo, J.; Du, L.; Chen, X.; Ma, X.; Fu, Q.; Han, S.; Zhang, D.; and Zhang, Y. 2023. On Manipulating Signals of User-Item Graph: A Jacobi Polynomial-based Graph Collaborative Filtering. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 602–613.

Guo, Q.; Zhuang, F.; Qin, C.; Zhu, H.; Xie, X.; Xiong, H.; and He, Q. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3549–3568.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Han, J.; Song, Y.; Zhao, W. X.; Shi, S.; and Zhang, H. 2018. hyperdoc2vec: Distributed Representations of Hypertext Documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2384–2394. Association for Computational Linguistics.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *International World Wide Web Conference*, 173–182.

Kozlov, M. 2023. 'Disruptive' science has declined — and no one knows why. *Nature*.

Kreutz, C. K.; and Schenkel, R. 2022. Scientific paper recommendation systems: a literature review of recent publications. *International Journal on Digital Libraries*, 23(4): 335–369.

Kurdi, A. 2021. *The Effects of Herd Mentality on Behavior*. Ph.D. thesis, Houston Baptist University.

Le, Q.; and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196. PMLR.

Liu, J.; Shi, C.; Yang, C.; Lu, Z.; and Philip, S. Y. 2022. A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources. *AI Open*, 3: 40–57.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mobasher, B.; Burke, R.; Bhaumik, R.; and Williams, C. 2007. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology*, 7(4): 23.

Park, M.; Leahey, E.; and Funk, R. J. 2023. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942): 138–144.

Peng, S.; Liu, X.; Sugiyama, K.; and Mine, T. 2024. How powerful is graph filtering for recommendation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 2388–2399.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Conference on Uncertainty in Artificial Intelligence*, 452–461.

Rohatgi, S.; Qin, Y.; Aw, B.; Unnithan, N.; and Kan, M.-Y. 2023. The ACL OCL Corpus: Advancing Open Science in Computational Linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10348–10361. Association for Computational Linguistics.

Saier, T.; and Färber, M. 2020. unarXive: A Large Scholarly Data Set with Publications' Full-Text, Annotated In-Text Citations, and Links to Metadata. *Scientometrics*, 125(3): 3085–3108.

Shahid, A.; Afzal, M. T.; Abdar, M.; Basiri, M. E.; Zhou, X.; Yen, N. Y.; and Chang, J.-W. 2020. Insights into relevant knowledge extraction techniques: a comprehensive review. *The Journal of Supercomputing*, 76: 1695–1733.

Si, M.; and Li, Q. 2020. Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review*, 53: 291–319.

Son, J.; and Kim, S. B. 2017. Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications*, 89: 404–412.

Su, X.-F.; Zeng, H.-J.; and Chen, Z. 2005. Finding group shilling in recommendation system. In *International World Wide Web Conference*, 960–961.

Sugiyama, K.; and Kan, M.-Y. 2010. Scholarly paper recommendation via user's recent research interests. In *Joint Conference on Digital Libraries (JCDL)*, 29–38.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. LINE: Large-scale Information Network Embedding. In *International World Wide Web Conference*. ACM.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. Arnetminer: extraction and mining of academic social networks. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 990–998.

West, J. D.; Wesley-Smith, I.; and Bergstrom, C. T. 2016. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, 2(2): 113–123.

Wu, L.; He, X.; Wang, X.; Zhang, K.; and Wang, M. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4425–4445.

Ye, H.; Li, X.; Yao, Y.; and Tong, H. 2023. Towards robust neural graph collaborative filtering via structure denoising and embedding perturbation. *ACM Transactions on Information Systems (TOIS)*, 41(3): 1–28.