MAGIC: TOPOLOGY-AWARE ANALYTIC GRAPH FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph few-shot class-incremental learning (GFSCIL) requires a model to continually recognize emerging classes from only a few labeled nodes while preserving previously acquired knowledge. Beyond the catastrophic forgetting inherited from conventional graph continual learning, GFSCIL presents two distinctive challenges: extremely limited novel-class supervision causes severe overfitting, while cross-session edges—edges connecting newly arriving nodes with historical nodes—alter historical propagation neighborhoods and thereby induce representation drift. We propose MAGIC, a replay-free GFSCIL framework that combines a frozen graph representation backbone (e.g., an intrinsically parameter-free backbone such as SGC or a pretrained graph foundation model) with closed-form analytic continual learning. To alleviate novel-session overfitting, MAGIC learns a topological prior from the base graph that can characterize both homophilous and heterophilous relations, and injects this prior through Potts Markov random field inference to refine supervision for novel classes. To mitigate representation drift, MAGIC transfers previous predictions from the old representations of affected historical nodes to their updated representations through drift-aware analytic distillation. Experiments across five datasets and eight baselines demonstrate the effectiveness of MAGIC. Under the 5-shot setting, MAGIC improves Mean Accuracy and Final Accuracy by 5.48 percentage points and 9.33 percentage points on average, and reduces Performance Drop by 10.78 percentage points on average compared with the best baselines. MAGIC also shows clear advantages under the 1- and 3-shot settings, with larger gains as the number of supports increases. Moreover, MAGIC requires substantially less training time.

1 INTRODUCTION

Graph neural networks (GNNs) have achieved remarkable success in node classification by integrating node attributes with relational information (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018). Most conventional GNNs, however, are developed under the assumption that the input graph and its label space remain static after training. This assumption rarely holds in real applications, where new entities and relations arrive continuously and the underlying graph evolves over time. Continual graph learning (CGL) therefore aims to incorporate newly arriving graph data while retaining knowledge acquired from previous stages (Zhou & Cao, 2021; Liu et al., 2021; Zhang et al., 2022).

An especially challenging yet practical setting arises when novel node classes emerge with only a handful of labeled nodes, known as *graph few-shot class-incremental learning* (GFSCIL) (Tan et al., 2022; Lu et al., 2022). In GFSCIL, a model first learns a set of base classes from a relatively well-labeled base graph. It then receives a sequence of incremental sessions, each introducing new nodes, new edges, and disjoint novel classes, but only limited labeled support nodes for each novel class. After every session, the model must classify all classes observed so far without revisiting the complete labeled data of previous sessions. Beyond the catastrophic forgetting problem inherited from GCIL, GFSCIL introduces two distinctive challenges.

Challenge 1: Novel-class overfitting. Each novel class is represented by only a few labeled support nodes, which provide insufficient supervision for estimating a reliable decision boundary. The model can therefore overfit the limited supports.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Challenge 2: Topology-induced representation drift. Nodes arriving in a new session may form edges with historical nodes, thereby altering the propagation neighborhoods of historical nodes and causing their representations to drift.

To address **Challenge 1**, existing GFSCIL methods employ pseudo-incremental meta-training and hierarchical attention (Tan et al., 2022), geometrically constrained prototype learning (Lu et al., 2022), structured prototype memories with node–memory interaction (Li et al., 2024), or iterative prototype calibration using pseudo-labeled query nodes (Li et al., 2025). However, these mechanisms operate primarily in the node-feature space and no longer care about the graph after feature extraction. They therefore overlook the potential of graph topological patterns as a prior for alleviating novel-session overfitting.

Challenge 2 remains largely overlooked in GFSCIL. Geometer employs teacher–student distillation, but evaluates the teacher on the current-session graph, which already contains newly arrived cross-session edges that were absent when the teacher was learned (Lu et al., 2022). Mecoin freezes the GNN encoder, yet fixed parameters do not prevent historical node representations from drifting when cross-session edges alter their neighborhoods (Li et al., 2024). GFCIL observes that a frozen encoder may produce overlapping representations when novel nodes are densely connected to base nodes, but does not model the reverse effect whereby these connections alter historical node representations (Tan et al., 2022).TAP instead considers an inductive setting with inaccessible or incomplete topology and therefore does not encounter the same issue (Li et al., 2025).

Motivated by these observations, we propose **MAGIC** (**M**arkov **A**nalytic **G**raph-**I**ncremental **C**lassifier). MAGIC employs a frozen graph representation backbone to extract node features and employs analytic learning (Zhuang et al., 2022) to accumulate knowledge through closed-form updates, thereby mitigating catastrophic forgetting without replay. However, as mentioned in **Challenge 2**, a fixed encoder does not guarantee invariant historical representations, MAGIC therefore identifies the affected historical nodes and performs drift-aware analytic distillation (Hinton et al., 2015): predictions produced by the previous classifier on the old graph are transferred as soft targets to their updated representations. To alleviate overfitting in novel sessions, MAGIC further learns from the base session a topological prior capable of characterizing both homophilous and heterophilous relations. It then injects this prior into the model through Potts Markov random field inference (Wang et al., 2021; Wainwright & Jordan, 2008).

Our main contributions are summarized as follows:

- We propose **MAGIC**, a replay-free GFSCIL framework with analytic continual learning.
- We develop a method for estimating a topological prior and inject it through Potts Markov random field inference to alleviate novel-session overfitting.
- To address the representation changes caused by newly arriving cross-session edges, we introduce analytic distillation.
- Extensive experiments across five datasets and eight baselines demonstrate that MAGIC consistently outperforms existing methods under the 1-, 3-, and 5-shot settings, and achieves substantially greater training efficiency.

2 PRELIMINARIES

Graph Few-Shot Class-Incremental Learning

The GFSCIL scenario is constructed by partitioning a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{y})$ into a base session and a sequence of novel sessions, where $\mathcal{V}$ and $\mathcal{E}$ are the node and edge sets, $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d_x}$ is the node-attribute matrix, and $\mathbf{y}$ contains the node labels. Following the GFSCIL setting (Tan et al., 2022; Lu et al., 2022), the label space is partitioned into one base-class set and T mutually disjoint novel-class sets: $\mathcal{C} = \mathcal{C}_0 \,\dot\cup\, \mathcal{C}_1 \,\dot\cup\, \cdots \,\dot\cup\, \mathcal{C}_T$, with $\mathcal{C}_{\le t} = \bigcup_{s=0}^{t} \mathcal{C}_s$.

Let $b(v)$ denote the session in which node v arrives, with $b(v) = t$ if $y_v \in \mathcal{C}_t$. At session t, the visible graph is

$$\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, \mathbf{X}_t), \qquad \mathcal{V}_t = \{v \in \mathcal{V} : b(v) \le t\}, \qquad \mathcal{E}_t = \{(u, v) \in \mathcal{E} : \max(b(u), b(v)) \le t\}. \tag{1}$$

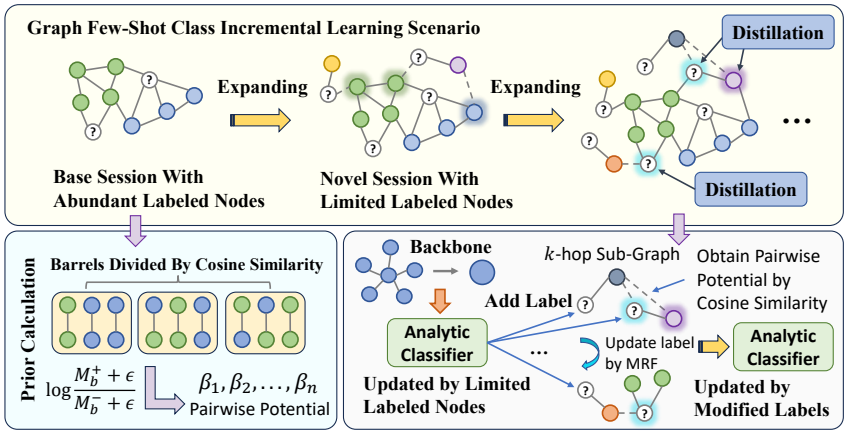


Figure 1: On the base stage, MAGIC learns a topological prior. In novel sessions, limited labeled supports first update an initial classifier, while drift-aware distillation protects historical nodes affected by cross-session edges. Finally, MAGIC extracts a subgraph around novel supports and performs Potts MRF inference with the topological prior to adjust the pseudo-labels predicted by the initial classifier, and subsequently updates the classifier using these adjusted labels.

In the base session, a subset $\mathcal{L}_c \subset \mathcal{V}_0$ is labeled for every $c \in \mathcal{C}_0$. At each novel session $t > 0$, every class $c \in \mathcal{C}_t$ provides a support set $\mathcal{S}_c \subset \mathcal{V}_t$ containing exactly K labeled nodes. The labels of all remaining nodes are unavailable. Historical labeled nodes cannot be stored or replayed after their supervision has been incorporated.

After session t, the learner produces a classifier $f_t : \mathcal{V}_t \to \mathcal{C}_{\leq t}$. Predictions are made over the complete set of classes encountered so far without a task identifier.

Frozen Graph Representation Backbones

MAGIC is compatible with a broad class of frozen graph representation backbones. These backbones may be intrinsically parameter-free graph propagation operators, such as SGC, Simple Spectral Graph Convolution (S^2GC), personalized PageRank diffusion and Heat Kernel (Wu et al., 2019; Zhu & Koniusz, 2021; Gasteiger et al., 2019; Xu et al., 2019), or pretrained graph foundation models used with frozen parameters, such as AnyGraph and RGVT (Xia & Huang, 2026; Lee et al., 2026).

3 Methodology

MAGIC is motivated by the following design rationale. It combines fixed graph encoder with analytic continual learning to preserve accumulated knowledge through closed-form updates and mitigate catastrophic forgetting without replay. Since cross-session edges can still alter historical node representations, MAGIC employs analytic distillation to transfer previous predictions to their updated representations. To alleviate novel-session overfitting, MAGIC estimates the relationship between the cosine similarity of connected node representations and their label agreement. As this relation is independent of specific class identities, it is more likely to generalize to novel classes. MAGIC then injects the resulting topological prior through MRF inference. The overall framework is illustrated in Figure 1.

3.1 Base Stage: Analytic Learning and Relational Prior

Let $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ be the base graph, $\mathcal{C}_0$ the set of base classes, and $\mathbf{h}_i \in \mathbb{R}^d$ the representation of node i. For class c, denote its labeled node set by $\mathcal{L}_c$ and its size by n_c.

Class-balanced analytic classifier. Directly assigning the same weight to every labeled node would make large base classes dominate the analytic solution. Moreover, each novel class sub-

sequently contributes unit support mass and can receive at most ρ additional pseudo-supervision mass (Further details are provided in Section 3.2.). MAGIC therefore assigns each base class total hard-label mass $1+\rho$, establishing the same maximum direct-supervision budget across base and novel classes. Accordingly, every labeled base node of class c receives weight $(1+\rho)/n_c$. The initial sufficient statistics and classifier are

$$\mathbf{R}_0 = \sum_{c\in\mathcal{C}_0}\sum_{i\in\mathcal{L}_c}\frac{1+\rho}{n_c}\mathbf{h}_i\mathbf{h}_i^\top, \qquad \mathbf{C}_0 = \sum_{c\in\mathcal{C}_0}\sum_{i\in\mathcal{L}_c}\frac{1+\rho}{n_c}\mathbf{h}_i\mathbf{e}_c^\top, \qquad \mathbf{W}_0 = (\mathbf{R}_0+\lambda\mathbf{I})^{-1}\mathbf{C}_0, \quad (2)$$

where $\mathbf{e}_c$ is the one-hot vector of class c, $\lambda > 0$ is the ridge coefficient, and $\rho \geq 0$ is the pseudo-supervision budget defined in Equation 12. The sufficient statistics $\mathbf{R}_0$ and $\mathbf{C}_0$ summarize the base supervision, so the original base nodes need not be replayed in subsequent sessions.

Class-balanced relation statistics. MAGIC further learns how a connected node pair's label agreement varies with the cosine similarity of its representations. We collect edges whose two endpoints are labeled base nodes and partition their cosine similarities into B quantile intervals $\{\mathcal{I}_1,\ldots,\mathcal{I}_B\}$. Quantile partitioning prevents the relation statistics from being concentrated in a small number of similarity intervals.

A naive edge count would again be biased toward large classes. For a same-class edge (i,j) from class c, we therefore assign the opportunity-normalized weight $w_{ij}^{+} = 1/(|\mathcal{C}_0|\binom{n_c}{2})$. For an edge connecting classes $c \neq c'$, we use $w_{ij}^{-} = 1/(\binom{|\mathcal{C}_0|}{2} n_c n_{c'})$. These weights are equivalent to first sampling a class or class pair uniformly and then sampling nodes within the selected classes.

For similarity interval $\mathcal{I}_b$, the balanced same-class and different-class edge masses are

$$M_b^{+} = \sum_{(i,j)\in\mathcal{E}_0} w_{ij}^{+}\mathbf{1}\{y_i = y_j\}\mathbf{1}\{\mathbf{h}_i^\top\mathbf{h}_j \in \mathcal{I}_b\}, \quad M_b^{-} = \sum_{(i,j)\in\mathcal{E}_0} w_{ij}^{-}\mathbf{1}\{y_i \neq y_j\}\mathbf{1}\{\mathbf{h}_i^\top\mathbf{h}_j \in \mathcal{I}_b\}. \quad (3)$$

The signed relation coefficient is then

$$\beta_b = \mathrm{clip}\left(\log\frac{M_b^{+}+\epsilon}{M_b^{-}+\epsilon}, -\beta_{\max}, \beta_{\max}\right), \quad (4)$$

where $\epsilon > 0$ stabilizes sparsely populated intervals.

A positive β_b favors equal labels, whereas a negative value favors different labels. MAGIC therefore does not impose a homophily assumption and can adapt to homophilous, heterophilous, or mixed graph relations. Moreover, β_b depends only on label agreement rather than class identity. It is consequently invariant to class permutation and is more likely to transfer from base classes to unseen classes.

3.2 Novel Stage: Analytic Learning and Topological Prior Injection

Step I: Naive Classifier with Drift Distillation

MAGIC first incorporates the labeled novel supports into the accumulated analytic statistics. Since each novel class provides exactly K supports, every support is assigned mass $1/K$. Before this update, zero columns for the newly introduced classes are appended to $\mathbf{C}_{t-1}$.

Although the graph backbone is either parameter-free or used with frozen parameters, newly arriving cross-session edges may change the propagation neighborhoods and thus the representations of historical nodes. Let $\mathbf{h}_{t-1,i}$ and $\mathbf{h}_{t,i}$ denote the previous and current representations of an affected historical node i, respectively. MAGIC measures its topology-induced drift as $\delta_i = \|\mathbf{h}_{t,i} - \mathbf{h}_{t-1,i}\|_2$ and assigns the normalized distillation mass

$$a_i^{D} = \kappa|\mathcal{C}_t^{\mathrm{new}}|\frac{\delta_i}{\sum_{j\in\mathcal{A}_t}\delta_j}, \qquad i \in \mathcal{A}_t, \quad (5)$$

where $\mathcal{A}_t$ is the set of affected historical nodes and κ determines the total distillation budget. The teacher target is obtained from the previous classifier and temperature-adjusted as

$$\mathbf{q}_i^{D} = \mathrm{Pad}_{|\mathcal{C}_{\leq t}|}\left[\mathrm{Temp}_{T_D}\left(\Pi_\Delta\left(\mathbf{h}_{t-1,i}^\top\mathbf{W}_{t-1}\right)\right)\right], \quad \mathrm{Temp}_{T_D}(\mathbf{p}) = \frac{\mathbf{p}^{1/T_D}}{\sum_c p_c^{1/T_D}}. \quad (6)$$

where Π_Δ projects scores onto the probability simplex and Pad appends zeros for newly introduced classes. The resulting soft class mass $\boldsymbol{\alpha}_i^D = a_i^D \mathbf{q}_i^D$ is imposed on the current representation by updating the analytic sufficient statistics:

$$\mathbf{R}_t \leftarrow \mathbf{R}_t + \sum_{i \in \mathcal{A}_t} a_i^D \mathbf{h}_{t,i} \mathbf{h}_{t,i}^\top, \quad \mathbf{C}_t \leftarrow \mathbf{C}_t + \sum_{i \in \mathcal{A}_t} \mathbf{h}_{t,i} (\boldsymbol{\alpha}_i^D)^\top. \tag{7}$$

Consequently, nodes undergoing larger topology-induced drift receive stronger protection while historical predictions are transferred without introducing artificial supervision for novel classes.

The influence of cross-session edges on analytic learning is summarized by the following theorem.

Theorem 1 (Topology-induced perturbation)**.** *Let $\Phi_{\boldsymbol{\theta}^\star}$ be a fixed graph encoder, and let $\mathbf{H}^-$ and $\mathbf{H}^+ = \mathbf{H}^- + \boldsymbol{\Delta}$ be the representations of a set of historical nodes before and after graph evolution. Given supervision matrix $\boldsymbol{\Gamma}$ and $\mathbf{D} = \mathrm{diag}(\boldsymbol{\Gamma}\mathbf{1})$, their analytic statistics satisfy*

$$\begin{aligned} \mathbf{R}^+ - \mathbf{R}^- &= (\mathbf{H}^-)^\top \mathbf{D}\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top \mathbf{D}\mathbf{H}^- + \boldsymbol{\Delta}^\top \mathbf{D}\boldsymbol{\Delta}, \\ \mathbf{C}^+ - \mathbf{C}^- &= \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}. \end{aligned} \tag{8}$$

Theorem 1 shows that a fixed graph encoder does not imply fixed analytic statistics on a dynamic graph, motivating the proposed drift-aware distillation. Its proof is provided in Appendix A.2.

After adding the class-balanced novel supports and distillation targets to the accumulated statistics, MAGIC obtains the naive classifier $\mathbf{W}_t^{(0)}$ through the same closed-form solution as Equation 2.

Step II: MRF Refinement and Analytic Reinjection

Starting from all novel-session supports, MAGIC extracts their k-hop induced subgraph from the currently visible graph. The subgraph can contain both historical and current-session nodes because the stage identity of an unlabeled node is not assumed to be known.

For every node i in this subgraph, the naive classifier produces a probability $\mathbf{p}_{0,i}$ over all classes observed so far, it is obtained using the same simplex projection as in Equation equation 6, but without temperature adjustment. This serves as the unary evidence. For a subgraph edge (i, j), MAGIC selects the base relation coefficient $\beta_{b(i,j)}$ according to the current cosine similarity and constructs the degree-normalized coupling $J_{ij} = \beta_{b(i,j)} / \sqrt{d_i d_j}$. The resulting Potts MRF is

$$p(\mathbf{y}) \propto \prod_i p_{0,i}(y_i)^{\tau_u} \prod_{(i,j) \in \mathcal{E}_t^{\mathrm{sub}}} \exp\left(\eta J_{ij} \mathbf{1}\{y_i = y_j\}\right), \tag{9}$$

where support nodes are clamped to their ground-truth labels. Degree normalization prevents hubs from accumulating excessive relational evidence.

MAGIC approximates the node marginals using damped mean-field inference:

$$\widehat{\mathbf{q}}_i^{(\ell+1)} = \mathrm{softmax}\left(\tau_u \log \mathbf{p}_{0,i} + \eta \sum_{j \in \mathcal{N}_i} J_{ij} \mathbf{q}_j^{(\ell)}\right), \qquad \mathbf{q}_i^{(\ell+1)} = (1-\alpha)\mathbf{q}_i^{(\ell)} + \alpha \widehat{\mathbf{q}}_i^{(\ell+1)}. \tag{10}$$

The support nodes are reset to their one-hot labels after every iteration.

Theorem 2 (Convergence of mean-field inference)**.** *Let $\mathbf{J}_{\mathrm{FF}}$ be the coupling matrix among free, unclamped nodes. If $\eta \|\mathbf{J}_{\mathrm{FF}}\|_2 / 2 < 1$, the damped mean-field mapping is a contraction under the Frobenius norm and converges to a unique fixed point. Its contraction factor is at most $1 - \alpha + \alpha\eta \|\mathbf{J}_{\mathrm{FF}}\|_2 / 2 < 1$.*

The proof is deferred to Appendix A.2. The complete probabilistic derivation is provided in Appendix A.1.

Novel-class posterior projection. The support-centered subgraph may contain both historical and current-session nodes. Labeled novel supports provide reliable class-specific anchors whose supervision can propagate to surrounding nodes through the MRF, whereas no labeled historical nodes are available to provide corresponding old-class anchors. MAGIC therefore uses the inferred posterior

to construct additional supervision for the current novel classes, rather than reinforcing historical-class predictions without labeled support.

Specifically, mean-field inference produces a calibrated posterior $\mathbf{q}_i$ over all classes observed up to session t. For every unlabeled candidate node, MAGIC retains its novel-class components without renormalization:

$$\widetilde{\gamma}_{i,c} = \begin{cases} q_i(c), & c \in \mathcal{C}_t^{\text{new}}, \\ 0, & c \notin \mathcal{C}_t^{\text{new}}. \end{cases} \tag{11}$$

This projection does not preselect or discard candidate nodes. Instead, the retained mass $\sum_{c \in \mathcal{C}_t^{\text{new}}} q_i(c)$ directly determines how strongly node i contributes to novel-class learning: nodes whose posterior is concentrated on historical classes naturally provide little novel-class supervision.

Conservative class balancing. The projected posteriors may assign substantially different total masses to different novel classes. Let $M_c = \sum_i \widetilde{\gamma}_{i,c}$ denote the available posterior mass for novel class c. MAGIC applies

$$\gamma_{i,c}^{\text{mrf}} = \min\left(1, \frac{\rho}{M_c + \epsilon}\right) \widetilde{\gamma}_{i,c}, \qquad c \in \mathcal{C}_t^{\text{new}}. \tag{12}$$

This limits each novel class to at most ρ pseudo-supervision mass while preserving the relative posterior allocation among candidate nodes. When the available evidence for a class is weaker than ρ, it remains unchanged rather than being artificially amplified.

Finally, let $\mathcal{U}_t$ denote the unlabeled candidate nodes and $m_i^{\text{mrf}} = \|\boldsymbol{\gamma}_i^{\text{mrf}}\|_1$ their supervision masses. The MRF-refined supervision is incorporated through

$$\mathbf{R}_t \leftarrow \mathbf{R}_t + \sum_{i \in \mathcal{U}_t} m_i^{\text{mrf}} \mathbf{h}_i \mathbf{h}_i^\top, \qquad \mathbf{C}_t \leftarrow \mathbf{C}_t + \sum_{i \in \mathcal{U}_t} \mathbf{h}_i (\boldsymbol{\gamma}_i^{\text{mrf}})^\top. \tag{13}$$

MAGIC then obtains $\mathbf{W}_t = (\mathbf{R}_t + \lambda \mathbf{I})^{-1} \mathbf{C}_t$. Thus, hard labels, distillation targets, and MRF-refined soft supervision are incorporated through the same analytic statistics.

4 Experiments

4.1 Setup

Datasets. We conduct experiments on five node-classification benchmarks: CoraFull (Bojchevski & Günnemann, 2018), Coauthor CS and Amazon Computers (Shchur et al., 2018), ogbn-arxiv (Hu et al., 2020), and WikiCS (Mernyei & Cangea, 2020).

Baselines. We compare MAGIC with general and graph continual-learning methods, including LwF (Li & Hoiem, 2016), EWC (Kirkpatrick et al., 2017), MAS (Aljundi et al., 2018), and TWP (Liu et al., 2021). We also compare against GFSCIL methods, including GFCIL (Tan et al., 2022), Geometer (Lu et al., 2022), Mecoin (Li et al., 2024), and TAP (Li et al., 2025).

Evaluation Metrics. We report three evaluation metrics. **MA (Mean Accuracy)** is the average overall accuracy across all sessions, and **FA (Final Accuracy)** is the overall accuracy after the final session. **PD** denotes the performance drop from the base session to the final session, with a smaller value indicating better retention. More details about datasets, baselines and backbones, and hyperparameter sensitivity analysis are shown in Appendix B

4.2 Main Results

Table 1 reports the 5-shot results on all five datasets.

MAGIC achieves the best MA, FA, and PD on CoraFull, ogbn-arxiv, Coauthor CS, and WikiCS. On Amazon Computers, it obtains the strongest FA and PD while remaining competitive in MA. Conventional continual learning methods deteriorate substantially because preserving previous parameters alone does not provide sufficient supervision for emerging classes, while GFSCIL-specific methods perform considerably better. Across graphs with different scales and structural characteristics, MAGIC consistently outperforms existing methods, demonstrating the effectiveness and robustness of the proposed framework.

Table 1: Main results under the 5-shot setting. MAGIC Backbone: SGC. MAGIC improves Mean Accuracy and Final Accuracy by 5.48 percentage points and 9.33 percentage points on average, and reduces Performance Drop by 10.78 percentage points on average compared with the best baselines.

	CoraFull			ogbn-arxiv			Coauthor CS			Amazon Computers			WikiCS		
Method	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD
LwF	16.06	3.82	79.86	18.16	4.65	82.03	35.83	17.16	79.73	35.05	13.55	82.07	35.83	12.90	79.78
LwF (std)	±1.24	±1.44	±2.35	±0.97	±5.80	±6.57	±3.44	±5.45	±5.73	±1.73	±5.61	±4.53	±2.33	±3.96	±4.33
EWC	16.14	3.86	79.82	15.87	6.07	80.52	36.38	17.32	79.58	35.13	13.96	81.54	35.89	12.86	79.82
EWC (std)	±1.31	±1.39	±2.30	±2.23	±6.70	±7.19	±3.85	±5.54	±5.78	±1.74	±5.92	±4.90	±2.49	±3.86	±4.27
MAS	16.31	6.07	77.61	34.40	27.37	59.25	45.91	20.77	76.12	34.98	13.35	82.17	35.24	12.42	80.26
MAS (std)	±1.21	±1.93	±2.13	±5.04	±10.00	±7.75	±5.15	±5.52	±5.51	±1.88	±5.41	±4.19	±2.04	±4.05	±4.50
TWP	17.48	5.05	78.21	32.36	17.24	68.38	39.70	15.93	80.97	34.76	13.37	82.45	35.80	13.24	78.94
TWP (std)	±1.49	±1.67	±3.38	±8.83	±5.84	±5.58	±5.54	±3.59	±3.78	±1.50	±5.54	±4.14	±1.82	±3.71	±3.52
GFCIL	54.23	41.33	39.81	56.77	40.78	42.02	78.02	71.29	24.95	**86.72**	80.01	12.61	76.63	67.18	23.74
GFCIL (std)	±1.64	±2.28	±5.31	±4.75	±3.34	±3.56	±3.98	±3.77	±5.35	±**2.79**	±2.20	±4.13	±2.77	±2.85	±3.71
Geometer	50.10	34.05	46.27	36.29	14.55	55.68	81.86	71.39	25.27	78.08	63.65	26.97	70.16	51.41	37.50
Geometer (std)	±1.74	±2.76	±4.73	±2.21	±3.51	±6.66	±3.49	±4.20	±3.21	±4.26	±4.16	±4.62	±4.74	±8.11	±8.20
Mecoin	47.57	32.60	44.13	46.20	31.08	46.18	71.81	54.16	42.07	70.44	56.31	32.13	65.17	50.12	38.62
Mecoin (std)	±1.60	±1.63	±4.54	±3.89	±0.89	±4.90	±4.00	±1.86	±2.33	±2.00	±3.86	±3.68	±3.97	±4.55	±4.65
TAP	59.27	45.06	38.48	57.97	42.69	42.37	79.81	67.83	28.76	83.76	74.06	21.10	76.24	63.66	28.16
TAP (std)	±1.35	±1.96	±4.04	±5.09	±1.23	±3.74	±1.87	±3.56	±4.04	±5.27	±4.90	±4.03	±3.27	±5.56	±6.78
MAGIC	**65.45**	**56.02**	**24.19**	**63.50**	**49.81**	**34.37**	**94.03**	**91.45**	**5.54**	86.54	**81.04**	**9.70**	**80.33**	**74.64**	**14.09**
MAGIC (std)	±**0.95**	±**1.05**	±**3.65**	±**3.92**	±**2.35**	±**2.55**	±**0.79**	±**0.28**	±**1.34**	±1.68	±**2.88**	±**5.06**	±**1.74**	±**2.14**	±**3.01**

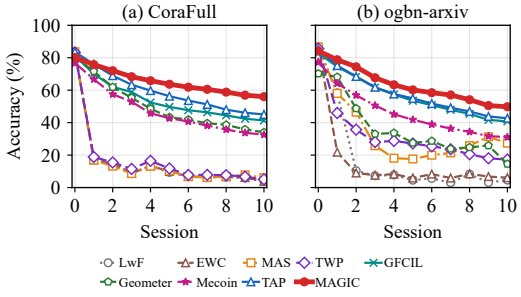


Figure 2: Overall accuracy across the base session and ten incremental sessions under the 5-shot setting. MAGIC Backbone: SGC

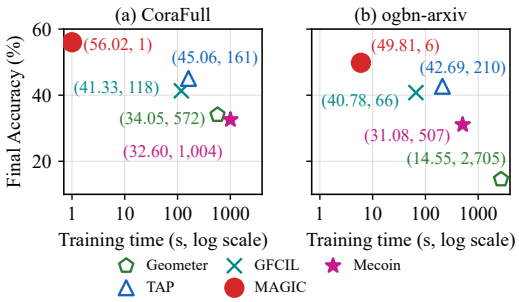


Figure 3: Final accuracy versus total training time under the 5-shot setting. MAGIC Backbone: SGC

Figure 2 presents the session-wise accuracy trajectories on CoraFull and ogbn-arxiv. MAGIC exhibits the most stable trajectory on both datasets, and its advantage becomes increasingly clear as novel classes accumulate. This indicate that MAGIC can more effectively acquire knowledge from sparsely supervised novel sessions while preserving previously acquired knowledge throughout the incremental process.

Figure 4 summarizes the 1- and 3-shot results on CoraFull. MAGIC achieves the strongest MA and FA and the smallest PD under both supervision levels. Taken together with the 5-shot results in Table 1, the advantage becomes more pronounced as additional supports are provided. Full results on 1- and 3-shot settings are provided in Appendix B.5.

4.3 Training Efficiency

We compare the total training time and final accuracy of GFSCIL methods on CoraFull and ogbn-arxiv. The results are shown in Figure 3.

MAGIC achieves the strongest final accuracy while requiring substantially less training time than competing GFSCIL methods on both datasets, demonstrating a favorable balance between predictive performance and computational efficiency.

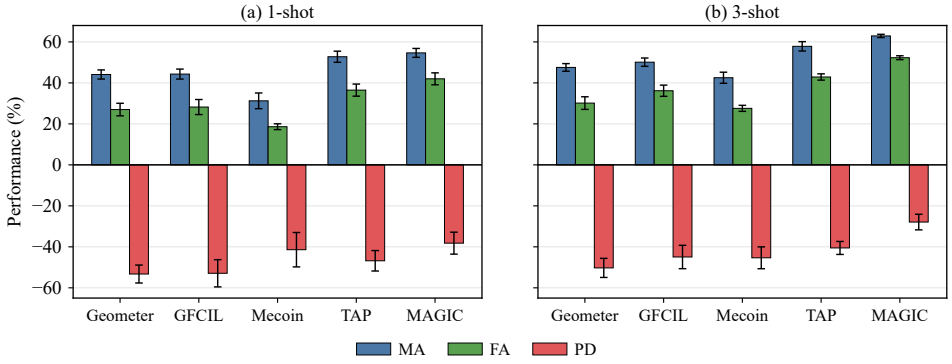


Figure 4: Main results on CoraFull under the 1- and 3-shot settings. MAGIC Backbone: SGC

Table 2: Results (%) with different frozen graph backbones under the 5-shot setting. RGVT is excluded from ogbn-arxiv because its released checkpoint was supervised-pretrained on that dataset.

	CoraFull			ogbn-arxiv			Coauthor CS			Amazon Computers			WikiCS		
Backbone	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD
SGC	65.45	56.02	24.19	63.50	49.81	34.37	94.03	91.45	5.54	86.54	81.04	**9.70**	80.33	74.64	**14.09**
SGC (std)	±0.95	±1.05	±3.65	±3.92	±2.35	±2.55	±0.79	±0.28	±1.34	±1.68	±2.88	**±5.06**	±1.74	±2.14	**±3.01**
S^2GC	64.42	54.74	25.48	61.23	47.05	36.43	94.25	91.77	5.43	**87.26**	**81.69**	9.76	**80.61**	**74.86**	14.20
S^2GC (std)	±0.57	±1.20	±4.24	±4.01	±2.39	±2.66	±0.78	±0.48	±1.40	**±1.58**	**±2.72**	±4.79	**±1.55**	**±2.09**	±3.34
PPR	**66.92**	**57.62**	**22.84**	**65.40**	**51.34**	**34.13**	94.26	91.90	5.02	85.89	79.96	10.58	80.12	73.90	14.98
PPR (std)	**±1.08**	**±0.51**	**±2.84**	**±4.13**	**±2.14**	**±2.52**	±0.63	±0.24	±1.42	±2.08	±2.49	±4.90	±1.98	±2.61	±3.58
Heat Kernel	57.96	46.71	31.06	53.99	39.30	40.32	93.81	90.91	6.62	85.56	78.55	12.66	79.48	73.49	14.80
Heat Kernel (std)	±0.72	±1.27	±4.44	±4.41	±3.29	±3.41	±0.90	±0.96	±1.66	±1.53	±2.72	±4.66	±1.53	±2.06	±2.95
RGVT	62.98	52.74	26.97	–	–	–	**94.57**	**92.19**	**4.99**	85.83	79.34	11.73	80.09	74.26	14.54
RGVT (std)	±0.43	±1.13	±4.20	–	–	–	**±0.53**	**±0.43**	**±1.54**	±1.97	±2.56	±4.50	±1.62	±2.32	±3.40

4.4 Effect of Frozen Graph Backbones

We evaluate MAGIC with five frozen graph representation backbones under the 5-shot setting. All other components and experimental protocols remain unchanged.

Table 2 shows that MAGIC is compatible with both parameter-free propagation operators and a frozen pretrained graph encoder. No single backbone dominates every graph: PPR performs best on the two citation networks, RGVT is strongest on Coauthor CS, and S^2GC achieves the highest accuracy on Amazon Computers and WikiCS, whereas SGC provides the smallest PD. These results demonstrate that MAGIC can benefit from backbone choices adapted to different graph characteristics.

4.5 Ablation Study

We ablate the two main components of MAGIC on ogbn-arxiv under the 5-shot setting. *w/o Prior* removes MRF-based topology-prior injection, and *w/o Distill* disables drift-aware analytic distillation. Results are shown in Table 3.

Table 3: Ablation results (%) on ogbn-arxiv under the 5-shot setting using SGC.

Variant	Mean Acc.	Final Acc.	PD
w/o Prior	53.69±7.74	37.29±7.28	46.89±6.72
w/o Distill	61.88±3.61	48.02±1.69	36.16±4.61
MAGIC	**63.50±3.92**	**49.81±2.35**	**34.37±2.55**

Removing the topology prior causes the largest drop, confirming that the topological prior provides complementary structural supervision for novel classes and thereby mitigates overfitting to the limited labeled supports. Disabling distillation also degrades overall accuracy and increases PD, showing that compensating for topology-induced representation drift improves historical retention. The complete model achieves the strongest results on all three metrics.

4.6 Novel Session Performance

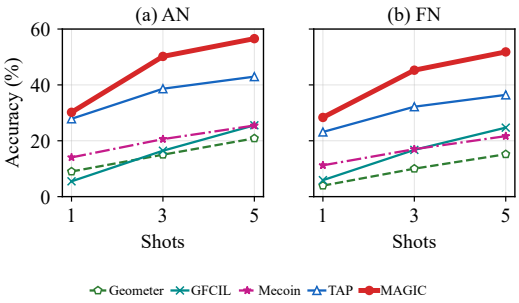


Figure 5: Novel-class performance on CoraFull.

Figure 5 shows novel-class performance under the 1-, 3-, and 5-shot settings on cora-full. AN is the average novel-class accuracy across incremental sessions, while FN measures the accuracy of all encountered novel classes after the final session. Both metrics generally improve as additional supports become available. MAGIC consistently achieves the strongest performance, with its advantage becoming more pronounced under the 3- and 5-shot settings. This demonstrates its ability to exploit limited supervision while retaining novel-class knowledge throughout the incremental process. More data about Novel Session Performance are provided in Appendix B.6.

5 Related Work

Continual Graph Learning Graph representation learning has evolved through both trainable architectures and fixed propagation schemes. GCN, GraphSAGE, and GAT learn task-specific encoder parameters (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018), whereas SGC, S^2GC, personalized PageRank, and graph heat kernels provide parameter-free graph propagation (Wu et al., 2019; Zhu & Koniusz, 2021; Gasteiger et al., 2019; Xu et al., 2019). More recently, graph foundation models such as AnyGraph and RGVT have explored transferable representations across diverse graphs (Xia & Huang, 2026; Lee et al., 2026). These advances nevertheless primarily consider static learning rather than sequential knowledge acquisition on an evolving graph. Continual graph learning addresses this setting by incorporating newly arriving graph data while preserving previous knowledge. ER-GNN employs experience replay (Zhou & Cao, 2021); CaT and PUMA compress historical information through graph condensation (Liu et al., 2023; 2025); DeLoMe develops debiased memory replay (Niu et al., 2024); TWP regularizes topology-sensitive parameters (Liu et al., 2021); and DyGRAIN distills knowledge for structurally affected nodes (Kim et al., 2022). CGLB provides standardized evaluation protocols for this setting (Zhang et al., 2022).

Graph Few-Shot Class-Incremental Learning GFSCIL extends few-shot class-incremental learning (Tao et al., 2020; Zhang et al., 2021; Zhou et al., 2022; Chi et al., 2022) to evolving graphs. HAG-Meta employs pseudo-incremental meta-training and hierarchical attention (Tan et al., 2022); Geometer learns geometrically constrained prototypes (Lu et al., 2022); Mecoin maintains structured prototype memories that interact with a GNN (Li et al., 2024); and TAP iteratively calibrates novel-class prototypes using pseudo-labeled query nodes in the inductive setting (Li et al., 2025).

Analytic Continual Learning Analytic continual learning replaces gradient-based classifier updates with closed-form solutions based on accumulated sufficient statistics. ACIL enables replay-free analytic class-incremental learning (Zhuang et al., 2022), while GKEAL extends this paradigm to few-shot incremental learning through Gaussian kernel embeddings (Zhuang et al., 2023). In the graph domain, GraphKeeper employs a recursively updated ridge-regression classifier to preserve decision boundaries across domains in graph domain-incremental learning (Guo et al., 2025).

6 Conclusion

We presented MAGIC, a replay-free framework for graph few-shot class-incremental learning. MAGIC combines frozen graph representation backbone, analytic continual learning, drift-aware distillation, and MRF-based topology-prior injection to address catastrophic forgetting, topology-induced representation drift, and novel-class overfitting. Extensive experiments demonstrate that MAGIC achieves strong accuracy and training efficiency across multiple GFSCIL benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and generating synthetic data sets is not applicable to this work. Additionally, we used generative AI tools for drafting parts of the research paper, summarizing or analysing existing literature, editing the paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. Specifically, all mathematical claims, proof steps, and derivations suggested or assisted by AI were independently verified line-by-line by the authors for correctness. Furthermore, all AI-identified or summarized references were manually cross-checked against the original literature to prevent hallucinated citations, and all translated/edited text was thoroughly reviewed for clarity, accuracy, and tone. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several measures to ensure the reproducibility of this work. The GFSCIL setting is provided in Section 2, analytic update rules, relational prior, MRF inference procedure, and drift distillation mechanism are specified in Section 3. Complete proofs of the theoretical results, together with their probabilistic interpretation, are provided in Appendix A. Dataset statistics, evaluation metrics, and the main experimental protocol are described in Section 4. Implementation details, software environment, random seeds, and per-dataset hyperparameters are reported in Appendix D. Additional results, including additinal main results, and session-wise novel-class performance, hyperparameter sensitivity analyses, are included in Appendix B. Anonymous source code sufficient to reproduce all reported experiments is submitted as supplementary material.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, pp. 139–154, 2018.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018.

Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. MetaFSCIL: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14166–14175, 2022.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22, 1977.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

Zihao Guo, Qingyun Sun, Ziwei Zhang, Haonan Yuan, Huiping Zhuang, Xingcheng Fu, and Jianxin Li. GraphKeeper: Graph domain-incremental learning via knowledge disentanglement and preservation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

Seoyoon Kim, Seongjun Yun, and Jaewoo Kang. DyGRAIN: An incremental learning framework for dynamic graphs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3157–3163, 2022. doi: 10.24963/ijcai.2022/438.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Dooho Lee, Myeong Kong, Minho Jeong, and Jaemin Yoo. View space: Learning representation across arbitrary graphs. In *International Conference on Machine Learning*, 2026.

Dong Li, Aijia Zhang, Junqi Gao, and Biqing Qi. An efficient memory module for graph few-shot class-incremental learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yayong Li, Peyman Moghadam, Can Peng, Nan Ye, and Piotr Koniusz. Inductive graph few-shot class incremental learning. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 466–474, 2025. doi: 10.1145/3701551.3703578.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629. Springer, 2016. doi: 10.1007/978-3-319-46493-0_37.

Huihui Liu, Yiding Yang, and Xinchao Wang. Overcoming catastrophic forgetting in graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8653–8661, 2021. doi: 10.1609/aaai.v35i10.17049.

Yilun Liu, Ruihong Qiu, and Zi Huang. CaT: Balanced continual graph learning with graph condensation. In *IEEE International Conference on Data Mining*, pp. 1157–1162, 2023. doi: 10.1109/ICDM58522.2023.00141.

Yilun Liu, Ruihong Qiu, Yanran Tang, Hongzhi Yin, and Zi Huang. PUMA: Efficient continual graph learning for node classification with graph condensation. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):449–461, 2025. doi: 10.1109/TKDE.2024.3485691.

Bin Lu, Xiaoying Gan, Lina Yang, Weinan Zhang, Luoyi Fu, and Xinbing Wang. Geometer: Graph few-shot class-incremental learning via prototype representation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1152–1161, 2022. doi: 10.1145/3534678.3539280.

Péter Mernyei and Cătălina Cangea. Wiki-CS: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Chaoxi Niu, Guansong Pang, and Ling Chen. Graph continual learning with debiased lossless memory replay. In *Proceedings of the 27th European Conference on Artificial Intelligence*, volume 392, pp. 1808–1815. IOS Press, 2024. doi: 10.3233/FAIA240692.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Zhen Tan, Kaize Ding, Ruocheng Guo, and Huan Liu. Graph few-shot class-incremental learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 987–996, 2022. doi: 10.1145/3488560.3498455.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12183–12192, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Binghui Wang, Jinyuan Jia, and Neil Zhenqiang Gong. Semi-supervised node classification on graphs: Markov random fields vs. graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):10093–10101, 2021. doi: 10.1609/aaai.v35i11.17211.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6861–6871. PMLR, 2019.

Lianghao Xia and Chao Huang. AnyGraph: Graph foundation model in the wild. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 882–896. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.44.

Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, and Xueqi Cheng. Graph convolutional networks using heat kernel for semi-supervised learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1928–1934. International Joint Conferences on Artificial Intelligence Organization, 2019. doi: 10.24963/ijcai.2019/267.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2021.

Xikun Zhang, Dongjin Song, and Dacheng Tao. CGLB: Benchmark tasks for continual graph learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 35, 2022.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9046–9056, 2022.

Fan Zhou and Chengtai Cao. Overcoming catastrophic forgetting in graph neural networks with experience replay. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4714–4722, 2021. doi: 10.1609/aaai.v35i5.16602.

Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11602–11614, 2022.

Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. GKEAL: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755, 2023. doi: 10.1109/CVPR52729.2023.00748.

A Theoretical Analysis

This appendix provides a probabilistic interpretation of the relational prior, Potts MRF inference, novel-class posterior projection, conservative class balancing, and analytic reinjection used by MAGIC. We then prove the topology-induced perturbation result and the convergence of damped mean-field inference.

A.1 Probabilistic Interpretation

A.1.1 The Base Relation Prior as a Balanced Log Bayes Factor

Let $S_{ij} = \mathbf{1}\{y_i = y_j\}$ indicate whether two labeled base nodes share the same class. For similarity interval $\mathcal{I}_b$, let $R_{ij}^{(b)}$ denote the event that (i, j) is connected and $\mathbf{h}_i^\top \mathbf{h}_j \in \mathcal{I}_b$. Therefore, $R_{ij}^{(b)}$ jointly describes graph connectivity and representation similarity, rather than similarity alone.

The raw number of same-class edges can be dominated by classes with many labeled nodes. MAGIC removes this dependence through a balanced pair-sampling interpretation. To sample a same-class pair, we first select a base class uniformly and then select an unordered pair uniformly from that class. Consequently, a pair from class c has probability $1/(|\mathcal{C}_0|\binom{n_c}{2})$, which is exactly the same-class weight w_{ij}^+ used in the base-stage statistics.

Similarly, to sample a different-class pair, we first select an unordered pair of distinct classes uniformly and then select one node uniformly from each class. A pair from classes c and c' therefore has probability $1/(\binom{|\mathcal{C}_0|}{2} n_c n_{c'})$, corresponding to w_{ij}^-.

Under these two balanced sampling distributions, define $\pi_b^+ = P_{\mathrm{bal}}(R_{ij}^{(b)} \mid S_{ij} = 1)$ and $\pi_b^- = P_{\mathrm{bal}}(R_{ij}^{(b)} \mid S_{ij} = 0)$. The weighted masses M_b^+ and M_b^- in Section 3.1 are empirical estimates of π_b^+ and π_b^-, respectively. The population relation coefficient is therefore

$$\beta_b^\star = \log \frac{P_{\mathrm{bal}}(R_{ij}^{(b)} \mid S_{ij} = 1)}{P_{\mathrm{bal}}(R_{ij}^{(b)} \mid S_{ij} = 0)} = \log \frac{\pi_b^+}{\pi_b^-}. \tag{14}$$

By Bayes' rule, the posterior odds of label agreement satisfy

$$\log \frac{P_{\mathrm{bal}}(S_{ij} = 1 \mid R_{ij}^{(b)})}{P_{\mathrm{bal}}(S_{ij} = 0 \mid R_{ij}^{(b)})} = \log \frac{P_{\mathrm{bal}}(S_{ij} = 1)}{P_{\mathrm{bal}}(S_{ij} = 0)} + \beta_b^\star. \tag{15}$$

Thus, $\beta_b^\star$ measures how much observing an edge in similarity interval $\mathcal{I}_b$ changes the log-odds that its endpoints have the same label. Equation 4 is a smoothed and clipped empirical estimate of this quantity. The smoothing constant prevents an undefined ratio when one type of edge is absent, while clipping limits the influence of unreliable estimates obtained from sparsely populated intervals.

To derive the corresponding pairwise potential, suppose the relational event depends on a pair of labels only through whether they agree. That is, $P(R_{ij}^{(b)} \mid y_i = a, y_j = c) = \pi_b^+$ when $a = c$, and equals π_b^- otherwise. This likelihood can be rewritten as

$$P(R_{ij}^{(b)} \mid y_i = a, y_j = c) = \pi_b^- \exp\left(\beta_b^\star \mathbf{1}\{a = c\}\right). \tag{16}$$

The factor π_b^- is independent of the candidate labels and is absorbed into the normalization constant. The remaining term is precisely a Potts pairwise potential.

When $\beta_b^\star > 0$, the observed relation is more likely under label agreement and the potential favors equal labels. When $\beta_b^\star < 0$, the relation is more likely under label disagreement and the potential discourages equal labels. Hence, the signed prior can express both homophilous and heterophilous relations.

For any permutation σ of the class identities, $\mathbf{1}\{y_i = y_j\} = \mathbf{1}\{\sigma(y_i) = \sigma(y_j)\}$. The prior therefore remains unchanged under class permutation. Its transfer from base to novel classes relies on the relation-stability assumption $P_t(R^{(b)} \mid S) = P_0(R^{(b)} \mid S)$: the feature distributions or semantic identities of the classes may change, but the conditional relationship between graph relations and label agreement should remain approximately stable.

A.1.2 From Analytic Scores to Probabilities

The analytic classifier produces an unconstrained score vector $\mathbf{s}_i = \mathbf{h}_i^\top \mathbf{W}$, which does not necessarily lie on the probability simplex. MAGIC converts it into a valid probability vector using Euclidean

simplex projection:

$$\Pi_\Delta(\mathbf{s}_i) = \arg\min_{\mathbf{p}\in\Delta^{C-1}} \frac{1}{2}\|\mathbf{p}-\mathbf{s}_i\|_2^2,$$
$$\Delta^{C-1} = \left\{\mathbf{p}\in\mathbb{R}^C_{\geq 0} : \mathbf{1}^\top\mathbf{p} = 1\right\}, \qquad \mathbf{p}_{0,i} = \Pi_\Delta\left(\mathbf{h}_{t,i}^\top \mathbf{W}_t^{(0)}\right). \tag{17}$$

The projection is the unique probability vector closest to the unconstrained analytic scores under the squared Euclidean geometry used by the ridge-regression objective. In implementation, its entries are lower-bounded by a small positive constant and renormalized to avoid undefined logarithms during MRF inference.

A.1.3 The Potts MRF Posterior

Let $\mathbf{p}_{0,i}$ be the full-class probability obtained from the naive analytic classifier through Equation 17. MAGIC defines the unary potential as $\phi_i(c) = p_{0,i}(c)^{\tau_u}$, where τ_u controls the confidence placed in the naive classifier. For a labeled support node, the unary potential is replaced by $\phi_i(c) = \mathbf{1}\{c = y_i\}$, thereby clamping its posterior to the observed label.

For edge (i,j), let $J_{ij} = \beta_{b(i,j)}/\sqrt{d_i d_j}$. The pairwise potential is $\psi_{ij}(a,c) = \exp(\eta J_{ij}\mathbf{1}\{a=c\})$, where η controls the global strength of the transferred prior. Combining all unary and pairwise potentials gives

$$p(\mathbf{y} \mid \mathcal{G}_t, \mathbf{H}_t, \mathcal{S}_t) = \frac{1}{Z}\prod_i p_{0,i}(y_i)^{\tau_u} \prod_{(i,j)\in\mathcal{E}_t^{\text{sub}}} \exp\left(\eta J_{ij}\mathbf{1}\{y_i = y_j\}\right), \tag{18}$$

which is the normalized form of Equation 9. Here, Z is the partition function obtained by summing the unnormalized probability over all assignments of the unclamped labels.

Without degree normalization and global scaling, the ratio between the equal-label and different-label pairwise potentials would be $\exp(\beta_b)$, corresponding directly to the estimated Bayes factor. With normalization, this ratio becomes $\exp(\eta\beta_b/\sqrt{d_i d_j})$. Thus, MAGIC can be interpreted as using a tempered or power likelihood: the direction of the relational evidence is preserved, while its magnitude is adjusted to prevent high-degree nodes from contributing the same evidence repeatedly through many incident edges.

The unary and pairwise evidence have distinct roles. The unary term specifies which particular class is plausible according to the naive classifier. The pairwise term does not contain a class-specific prototype and only evaluates whether neighboring labels should agree. The MRF posterior therefore combines class-specific predictions with a class-permutation-invariant relation model.

A.1.4 Mean-Field Inference as Variational Optimization

Exact marginalization of Equation 18 is generally intractable. MAGIC uses the factorized approximation $q(\mathbf{y}) = \prod_{i\in\mathcal{V}_{\text{free}}} q_i(y_i)$, while the support-node distributions remain fixed one-hot vectors.

Mean-field inference minimizes $\mathrm{KL}(q(\mathbf{y})\|p(\mathbf{y}))$. After discarding terms that do not depend on q, the variational free energy is

$$\mathcal{F}(q) = \sum_{i\in\mathcal{V}_{\text{free}}}\sum_c q_i(c)\log q_i(c) - \tau_u \sum_{i\in\mathcal{V}_{\text{free}}}\sum_c q_i(c)\log p_{0,i}(c)$$
$$- \eta \sum_{(i,j)\in\mathcal{E}_t^{\text{sub}}} J_{ij}\sum_c q_i(c)q_j(c). \tag{19}$$

The first term is the negative entropy of the variational distribution. The second encourages agreement with the naive classifier, while the third encourages or discourages neighboring label agreement according to the sign of J_{ij}.

To minimize $\mathcal{F}$ with respect to q_i, introduce a Lagrange multiplier λ_i for the constraint $\sum_c q_i(c) = 1$. The partial derivative is $\partial\mathcal{F}/\partial q_i(c) = \log q_i(c) + 1 - \tau_u \log p_{0,i}(c) - \eta\sum_{j\in\mathcal{N}_i} J_{ij}q_j(c) + \lambda_i$. Setting it to zero gives $\log q_i(c) = \tau_u \log p_{0,i}(c) + \eta\sum_{j\in\mathcal{N}_i} J_{ij}q_j(c) - 1 - \lambda_i$.

Exponentiating this equality and normalizing over the classes yields

$$q_i(c) = \frac{p_{0,i}(c)^{\tau_u} \exp\left(\eta \sum_{j \in \mathcal{N}_i} J_{ij} q_j(c)\right)}{\sum_{c'} p_{0,i}(c')^{\tau_u} \exp\left(\eta \sum_{j \in \mathcal{N}_i} J_{ij} q_j(c')\right)}. \tag{20}$$

This is equivalent to the softmax proposal in Equation 10. Damping replaces the proposal $\widehat{\mathbf{q}}_i$ by the convex combination $(1-\alpha)\mathbf{q}_i + \alpha\widehat{\mathbf{q}}_i$. Since $\alpha > 0$, a distribution is a fixed point of the damped update if and only if it is a fixed point of the original mean-field update.

A.1.5 An EM-Like Interpretation of Novel-Stage Learning

The two-step novel-stage update follows an EM-like predict–refine–update procedure (Dempster et al., 1977). Given the few-shot supports and distillation targets, MAGIC first obtains the naive classifier $\mathbf{W}_t^0$. Analogous to an E-step, it then treats the labels of unlabeled candidate nodes as latent variables and uses Mean-Field inference to estimate their graph-refined posterior distributions $\mathbf{q}_i$. Analogous to an M-step, these posteriors are converted into supervision $\boldsymbol{\alpha}_i$ and inserted into the analytic sufficient statistics, after which $\mathbf{W}_t^1$ is obtained in closed form. The overall update can therefore be summarized as $\mathbf{W}_t^0 \to \{\mathbf{q}_i\} \to \{\boldsymbol{\alpha}_i\} \to \mathbf{W}_t^1$.

A.1.6 Unified Probabilistic Interpretation of Analytic Injection

Let $\boldsymbol{\gamma}_i \in \mathbb{R}^C_{\geq 0}$ be any nonnegative supervision vector. It may represent a balanced hard label, a distillation target, or an MRF pseudo-label. Define its total mass as $m_i = \|\boldsymbol{\gamma}_i\|_1$. If $m_i > 0$, its normalized target is $\mathbf{p}_i = \boldsymbol{\gamma}_i / m_i$.

Consider the Gaussian working likelihood

$$\mathbf{p}_i \mid \mathbf{h}_i, \mathbf{W}, m_i \sim \mathcal{N}\left(\mathbf{h}_i^\top \mathbf{W}, \frac{1}{2m_i}\mathbf{I}\right). \tag{21}$$

A larger supervision mass corresponds to lower observation variance and therefore greater confidence. Assume the Gaussian parameter prior $p(\mathbf{W}) \propto \exp(-\lambda\|\mathbf{W}\|_F^2)$. Ignoring constants independent of $\mathbf{W}$, the negative log-posterior becomes

$$\mathcal{L}(\mathbf{W}) = \sum_i m_i \left\|\mathbf{h}_i^\top \mathbf{W} - \mathbf{p}_i\right\|_2^2 + \lambda\|\mathbf{W}\|_F^2. \tag{22}$$

Since $m_i \mathbf{p}_i = \boldsymbol{\gamma}_i$, differentiating with respect to $\mathbf{W}$ and setting the derivative to zero gives

$$\left(\sum_i m_i \mathbf{h}_i \mathbf{h}_i^\top + \lambda \mathbf{I}\right) \mathbf{W} = \sum_i \mathbf{h}_i \boldsymbol{\gamma}_i^\top. \tag{23}$$

Thus, defining $\mathbf{R} = \sum_i m_i \mathbf{h}_i \mathbf{h}_i^\top$ and $\mathbf{C} = \sum_i \mathbf{h}_i \boldsymbol{\gamma}_i^\top$ recovers $\mathbf{W} = (\mathbf{R} + \lambda \mathbf{I})^{-1}\mathbf{C}$, consistent with Equation 2.

This result explains why the three supervision sources are mathematically compatible. A true one-hot label is a probability target with explicit class-balanced precision. A distillation distribution is a soft observation that preserves the historical decision function. An MRF posterior is a graph-refined soft observation whose supervision mass is determined by novel-class posterior projection and conservative class balancing. The retained novel-class mass controls each node's contribution, while class-level scaling limits the aggregate pseudo-supervision mass to at most ρ per novel class.

A.2 Proofs of Theoretical Results

A.2.1 Proof of Theorem 1

Let $\mathbf{H}^+ = \mathbf{H}^- + \boldsymbol{\Delta}$. Before graph evolution, the historical nodes contribute $\mathbf{R}^- = (\mathbf{H}^-)^\top \mathbf{D} \mathbf{H}^-$ and $\mathbf{C}^- = (\mathbf{H}^-)^\top \boldsymbol{\Gamma}$. After graph evolution,

$$\begin{aligned} \mathbf{R}^+ &= (\mathbf{H}^- + \boldsymbol{\Delta})^\top \mathbf{D} (\mathbf{H}^- + \boldsymbol{\Delta}) \\ &= \mathbf{R}^- + (\mathbf{H}^-)^\top \mathbf{D}\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top \mathbf{D}\mathbf{H}^- + \boldsymbol{\Delta}^\top \mathbf{D}\boldsymbol{\Delta}, \end{aligned} \tag{24}$$

and $\mathbf{C}^+ = (\mathbf{H}^- + \boldsymbol{\Delta})^\top \boldsymbol{\Gamma} = \mathbf{C}^- + \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}$. Subtracting the previous statistics proves both identities. □

A.2.2 Proof of Theorem 2

Only free nodes need to be considered because support-node distributions are clamped. Messages from support nodes are constant and can therefore be absorbed into a unary matrix $\mathbf{U}$. Let $\mathbf{Q} \in \mathbb{R}^{n_f \times C}$ contain the free-node distributions and let $\mathbf{J}_{\mathrm{FF}}$ be their coupling matrix. The undamped mean-field mapping is

$$\mathcal{T}(\mathbf{Q}) = \mathrm{Softmax}_{\mathrm{row}}\left(\mathbf{U} + \eta \mathbf{J}_{\mathrm{FF}}\mathbf{Q}\right). \tag{25}$$

We first derive a Lipschitz bound for softmax. For $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$, its Jacobian is $\mathbf{J}_{\mathrm{sm}}(\mathbf{p}) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$. This matrix is symmetric and positive semidefinite. In row c, the diagonal entry is $p_c(1-p_c)$ and the sum of the absolute off-diagonal entries is also $p_c(1-p_c)$.

By the Gershgorin circle theorem, every eigenvalue of $\mathbf{J}_{\mathrm{sm}}$ is no larger than $2p_c(1-p_c)$ for some c. Since $p_c(1-p_c) \le 1/4$, it follows that $\|\mathbf{J}_{\mathrm{sm}}(\mathbf{p})\|_2 \le 1/2$. The mean-value theorem therefore gives

$$\|\,\mathrm{softmax}(\mathbf{z}) - \mathrm{softmax}(\mathbf{z}')\|_2 \le \frac{1}{2}\|\mathbf{z} - \mathbf{z}'\|_2. \tag{26}$$

Applying Equation 26 independently to every row of Equation 25 yields

$$\begin{aligned}\|\mathcal{T}(\mathbf{Q}) - \mathcal{T}(\mathbf{Q}')\|_F &\le \frac{\eta}{2}\|\mathbf{J}_{\mathrm{FF}}(\mathbf{Q} - \mathbf{Q}')\|_F \\ &\le \frac{\eta}{2}\|\mathbf{J}_{\mathrm{FF}}\|_2\|\mathbf{Q} - \mathbf{Q}'\|_F.\end{aligned} \tag{27}$$

The damped mapping is $\mathcal{T}_\alpha(\mathbf{Q}) = (1-\alpha)\mathbf{Q} + \alpha\mathcal{T}(\mathbf{Q})$. Thus,

$$\begin{aligned}\|\mathcal{T}_\alpha(\mathbf{Q}) - \mathcal{T}_\alpha(\mathbf{Q}')\|_F \le&(1-\alpha)\|\mathbf{Q} - \mathbf{Q}'\|_F \\ &+ \alpha\|\mathcal{T}(\mathbf{Q}) - \mathcal{T}(\mathbf{Q}')\|_F \\ &\le \left(1 - \alpha + \frac{\alpha\eta}{2}\|\mathbf{J}_{\mathrm{FF}}\|_2\right)\|\mathbf{Q} - \mathbf{Q}'\|_F.\end{aligned} \tag{28}$$

Define $\kappa_{\mathrm{MF}} = 1 - \alpha + \alpha\eta\|\mathbf{J}_{\mathrm{FF}}\|_2/2$. Under the condition $\eta\|\mathbf{J}_{\mathrm{FF}}\|_2/2 < 1$, we have $\kappa_{\mathrm{MF}} < 1$, so the damped mapping is a contraction.

The product of node-wise probability simplices is a closed subset of a finite-dimensional Euclidean space and is therefore complete. By the Banach fixed-point theorem, the mapping has a unique fixed point $\mathbf{Q}^\star$. Repeatedly applying the contraction inequality gives

$$\|\mathbf{Q}^{(\ell)} - \mathbf{Q}^\star\|_F \le \kappa_{\mathrm{MF}}^\ell\|\mathbf{Q}^{(0)} - \mathbf{Q}^\star\|_F. \tag{29}$$

Therefore, the damped mean-field iterates converge geometrically to the unique fixed point, completing the proof of Theorem 2. □

B Further Experiments

B.1 Datasets

We evaluate MAGIC on five widely used node-classification benchmarks: CoraFull (Bojchevski & Günnemann, 2018), Coauthor CS (Shchur et al., 2018), ogbn-arxiv (Hu et al., 2020), WikiCS (Mernyei & Cangea, 2020), and Amazon Computers (Shchur et al., 2018). CoraFull and ogbn-arxiv are citation networks in which nodes represent papers and labels correspond to research topics. Coauthor CS connects authors who have coauthored papers, with labels indicating their primary research fields. WikiCS is a hyperlink network of computer-science articles from Wikipedia, while Amazon Computers is a product co-purchase network whose labels denote product categories. Table 4 summarizes the dataset statistics. We report edge homophily $h_{\mathrm{edge}} = |\{(u,v) \in \mathcal{E} : y_u = y_v\}|/|\mathcal{E}|$, which measures the proportion of edges connecting nodes with the same label.

B.2 Class-Incremental Splits

Table 5 reports the number of base classes and novel classes introduced per incremental session for each dataset.

Table 4: Statistics of the node-classification datasets.

Dataset	Nodes	Edges	Features	Classes	h_{edge}
CoraFull	19,793	126,842	8,710	70	0.57
Coauthor CS	18,333	163,788	6,805	15	0.81
ogbn-arxiv	169,343	1,166,243	128	40	0.65
WikiCS	11,701	216,123	300	10	0.65
Amazon Computers	13,752	491,722	767	10	0.78

Table 5: Class-incremental split configurations.

Dataset	Base Classes	Way
CoraFull	20	5
ogbn-arxiv	10	3
Coauthor CS	5	2
Amazon Computers	4	2
WikiCS	4	2

B.3 Compared Baselines

We compare MAGIC with four representative continual-learning methods and four GFSCIL-specific methods. All baselines are evaluated using the same class splits, support sets, visible graphs, and evaluation protocol as MAGIC.

LwF. Learning without Forgetting (LwF) preserves previous knowledge through knowledge distillation (Li & Hoiem, 2016). Before learning a new session, it records the predictions of the previous model and uses them as soft targets while optimizing the current model. It does not require storing historical samples, but its effectiveness depends on whether the newly available data can provide sufficient coverage of previously learned classes.

EWC. Elastic Weight Consolidation (EWC) estimates the importance of each model parameter using the diagonal of the Fisher information matrix (Kirkpatrick et al., 2017). During subsequent sessions, it adds a quadratic penalty that discourages important parameters from moving away from their previous values. EWC primarily protects historical knowledge at the parameter level and does not explicitly address the scarcity of novel-class supervision.

MAS. Memory Aware Synapses (MAS) measures parameter importance according to the sensitivity of the model output to changes in each parameter (Aljundi et al., 2018). Unlike EWC, its importance estimation does not require labels and can therefore exploit unlabeled inputs. Important parameters are subsequently protected by a weighted quadratic regularizer during incremental learning.

TWP. Topology-aware Weight Preserving (TWP) is a continual graph-learning method specifically designed for GNNs (Liu et al., 2021). In addition to measuring parameter importance for node classification, TWP identifies parameters important to neighborhood aggregation and topological information. It regularizes these parameters in later sessions to mitigate forgetting of both predictive and structural knowledge.

GFCIL. The original GFSCIL framework introduces graph pseudo-incremental learning and the HAG-Meta model (Tan et al., 2022). It constructs pseudo-incremental episodes from base classes to simulate future few-shot sessions and employs hierarchical attention at the task and node levels. A task-sensitive regularizer is further used to balance adaptation to novel classes and retention of base knowledge.

Geometer. Geometer represents each class through an attention-based prototype and imposes geometric constraints on the evolving prototype space (Lu et al., 2022). Its prototype representation incorporates graph information to improve class separability, while knowledge distillation helps

preserve previously learned classes. The resulting classifier performs incremental prediction by comparing node representations with the accumulated class prototypes.

Mecoin. Mecoin introduces a structured memory unit for efficiently storing and updating class prototypes (Li et al., 2024). Its memory-construction module enables interactions between incoming nodes and stored prototypes, while a memory-representation adaptation module separates prototype learning from class prediction. It further exchanges knowledge between the memory and GNN to reduce extensive encoder fine-tuning and preserve historical information.

TAP. TAP addresses inductive GFSCIL through topology-based class augmentation and prototype calibration (Li et al., 2025). It augments the limited novel-class supports using graph-topological information and iteratively calibrates novel prototypes with confident pseudo-labeled query nodes. This improves few-shot generalization while retaining a prototype-based incremental classifier.

B.4 Backbone Implementations

We evaluate five frozen graph representation backbones. For the four parameter-free operators, let $\mathbf{P}_t = \widetilde{\mathbf{D}}_t^{-1/2}\widetilde{\mathbf{A}}_t\widetilde{\mathbf{D}}_t^{-1/2}$ denote the symmetrically normalized adjacency matrix with self-loops, and let $\mathbf{X}'_t$ denote the node attributes after the optional fixed random projection. All resulting node representations are ℓ_2-normalized row-wise before being used by MAGIC.

SGC. Simplified Graph Convolution (SGC) removes intermediate nonlinearities and trainable transformations, retaining only fixed graph propagation followed by a linear classifier (Wu et al., 2019). Its representation is

$$\overline{\mathbf{H}}_t = \mathbf{P}_t^K \mathbf{X}'_t, \tag{30}$$

where K is both the propagation depth and the maximum receptive-field radius. SGC retains only the representation obtained after the final propagation step.

S^{2}GC. Simple Spectral Graph Convolution (S^{2}GC) combines representations from multiple propagation depths rather than retaining only the Kth-order result (Zhu & Koniusz, 2021). We implement

$$\overline{\mathbf{H}}_t = \alpha\mathbf{X}'_t + \frac{1-\alpha}{K}\sum_{\ell=1}^{K}\mathbf{P}_t^\ell\mathbf{X}'_t, \tag{31}$$

where $\alpha \in [0,1]$ controls the residual contribution of the original attributes. This operator averages, rather than concatenates, the different propagation orders and has a maximum range of K hops.

PPR. Personalized PageRank (PPR) propagation repeatedly diffuses node features while restarting from the original attributes (Gasteiger et al., 2019). Starting from $\mathbf{H}_t^{(0)} = \mathbf{X}'_t$, our finite-step implementation uses

$$\mathbf{H}_t^{(\ell+1)} = (1-\alpha)\mathbf{P}_t\mathbf{H}_t^{(\ell)} + \alpha\mathbf{X}'_t, \qquad \ell = 0,\dots,K-1, \tag{32}$$

and returns $\overline{\mathbf{H}}_t = \mathbf{H}_t^{(K)}$. Here, α is the restart probability. Unlike the exact infinite-step PPR solution, this truncated recurrence requires only K sparse propagations and restricts the influence range to at most K hops.

Truncated Heat Kernel. Heat-kernel diffusion propagates graph signals continuously through $\exp[-\tau(\mathbf{I}-\mathbf{P}_t)]$ (Xu et al., 2019). We avoid matrix exponentiation by truncating its Taylor expansion:

$$\overline{\mathbf{H}}_t = \sum_{\ell=0}^{K} w_\ell \mathbf{P}_t^\ell \mathbf{X}'_t, \qquad w_\ell = \frac{\tau^\ell/\ell!}{\sum_{r=0}^{K}\tau^r/r!}. \tag{33}$$

The diffusion time $\tau \geq 0$ determines how strongly higher-order neighborhoods contribute, while K controls the truncation order and strict maximum hop range. The common factor $\exp(-\tau)$ cancels under coefficient and row normalization.

Table 6: Main results (%) under the 1-shot setting. The mean and standard deviation over five runs are reported in consecutive rows. The best result in each column is shown in bold. MAGIC Backbone: SGC.

	CoraFull			ogbn-arxiv			Coauthor CS			Amazon Computers			WikiCS		
Method	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD
LwF	13.18	2.28	81.39	16.64	5.29	81.38	35.23	16.68	80.21	32.84	13.17	82.40	34.56	11.45	81.14
LwF (std)	±0.93	±0.93	±2.91	±1.53	±6.77	±7.35	±3.34	±5.40	±5.82	±2.10	±5.34	±4.25	±2.40	±4.66	±4.76
EWC	13.53	2.83	80.85	17.48	9.11	77.57	36.50	17.14	79.76	33.16	14.63	80.92	34.83	11.76	80.91
EWC (std)	±1.09	±1.71	±2.75	±1.25	±5.30	±5.67	±4.02	±5.25	±5.64	±2.62	±6.77	±5.92	±2.48	±4.98	±5.12
MAS	16.40	6.16	77.52	30.77	18.51	68.16	51.31	24.55	72.34	34.88	12.71	82.95	35.65	12.35	80.26
MAS (std)	±1.80	±2.48	±3.02	±7.41	±3.40	±6.17	±10.34	±9.14	±8.87	±1.64	±5.01	±4.10	±2.47	±3.94	±4.33
TWP	15.60	3.98	79.20	34.73	18.94	66.68	39.70	16.54	80.34	32.85	12.53	83.28	34.60	11.41	80.75
TWP (std)	±1.06	±1.31	±3.12	±9.08	±4.69	±3.58	±3.42	±8.27	±8.77	±1.13	±5.32	±4.23	±1.94	±4.92	±4.75
GFCIL	44.30	28.22	52.91	47.80	30.00	52.74	54.15	31.32	64.92	79.05	64.06	28.58	68.82	53.56	37.10
GFCIL (std)	±2.44	±3.67	±6.65	±8.56	±4.75	±3.59	±7.73	±6.10	±7.08	±3.42	±6.17	±8.39	±4.82	±6.47	±6.70
Geometer	44.10	26.99	53.27	36.31	19.71	50.50	64.60	42.82	53.79	73.09	58.54	32.13	62.49	41.91	47.06
Geometer (std)	±2.23	±3.06	±4.38	±5.59	±5.08	±9.03	±8.61	±13.23	±13.01	±4.10	±5.74	±3.51	±4.04	±7.48	±8.97
Mecoin	31.25	18.62	41.40	24.19	15.63	**31.48**	57.86	37.38	54.51	56.23	39.48	34.70	49.33	35.69	32.40
Mecoin (std)	±3.83	±1.42	±8.39	±4.29	±0.62	±**6.77**	±9.39	±7.60	±5.46	±8.20	±7.36	±8.82	±9.04	±8.03	±5.47
TAP	52.76	36.47	46.81	54.47	38.62	46.43	76.85	64.02	32.54	81.02	68.44	26.79	73.67	58.91	32.87
TAP (std)	±2.71	±2.95	±4.99	±5.82	±2.91	±2.27	±3.37	±6.66	±6.39	±6.33	±7.82	±7.16	±3.33	±4.76	±5.76
MAGIC	**54.64**	**41.99**	**38.21**	**55.48**	**40.71**	43.47	**89.97**	**86.48**	**10.51**	**84.06**	**76.68**	**14.06**	**76.82**	**65.68**	**23.05**
MAGIC (std)	±**2.17**	±**2.91**	±**5.36**	±**7.96**	±**5.37**	±3.90	±**3.27**	±**3.94**	±**4.84**	±**3.37**	±**5.67**	±**7.78**	±**2.88**	±**3.93**	±**3.95**

RGVT. Recurrent Graph View Transformation (RGVT) is a pretrained graph representation model that applies a shared scalar transformation to multi-hop graph views (Lee et al., 2026). Let $\mathbf{P}_t^{\mathrm{rw}}$ and $\mathbf{P}_t^{\mathrm{sym}}$ denote the random-walk and symmetric propagation operators, respectively. For node i and feature channel f, the released encoder constructs

$$\mathbf{v}_{i,f}(\mathbf{H}) = \left[H_{i,f}, (\mathbf{P}_t^{\mathrm{rw}}\mathbf{H})_{i,f}, (\mathbf{P}_t^{\mathrm{sym}}\mathbf{H})_{i,f}, ((\mathbf{P}_t^{\mathrm{rw}})^2\mathbf{H})_{i,f}, ((\mathbf{P}_t^{\mathrm{sym}})^2\mathbf{H})_{i,f}\right] \tag{34}$$

and computes $H_{i,f}^{(r+1)} = g_{\boldsymbol{\theta}}(\mathbf{v}_{i,f}(\mathbf{H}^{(r)}))$ using the same pretrained MLP $g_{\boldsymbol{\theta}}$ for every node and feature channel. MAGIC freezes $\boldsymbol{\theta}$ and applies this transformation recurrently for R rounds, producing an effective receptive field of at most $2R$ hops. The released checkpoint fixes the view order and MLP depth to two. We exclude RGVT from ogbn-arxiv because the checkpoint was supervised-pretrained on that dataset, which would cause information leakage.

B.5 Additional Main Results

Tables 6 and 7 report the complete 1- and 3-shot results on all five datasets. The corresponding 5-shot results are presented in the main text.

Across the 1- and 3-shot settings, MAGIC achieves the highest MA and FA on every dataset. It also obtains the smallest PD in all cases except 1-shot ogbn-arxiv, where Mecoin exhibits a smaller decline but substantially lower overall and final accuracy. The improvements are particularly consistent on Coauthor CS and WikiCS, while the results on the larger ogbn-arxiv graph demonstrate that MAGIC remains effective under substantially different graph scales and structural characteristics. Together with the 5-shot results in the main text, these experiments show that MAGIC provides a robust balance between novel-class adaptation and historical knowledge retention across different supervision levels.

B.6 Novel-Class Performance

We further evaluate how effectively each method learns and retains novel classes. AN and FN denote the average novel-class accuracy across incremental sessions and the novel-class accuracy after the final session, respectively. Tables 8– 10 report the results.

MAGIC achieves the strongest AN and FN in every 3- and 5-shot setting, demonstrating that it both learns novel classes effectively and preserves them as subsequent classes arrive. Under the more challenging 1-shot setting, MAGIC remains best on four of the five datasets, while TAP performs better on ogbn-arxiv. The advantages become particularly clear as additional supports are provided,

Table 7: Main results (%) under the 3-shot setting. The mean and standard deviation over five runs are reported in consecutive rows. The best result in each column is shown in bold. MAGIC Backbone: SGC.

	CoraFull			ogbn-arxiv			Coauthor CS			Amazon Computers			WikiCS		
Method	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD	MA	FA	PD
LwF	15.46	3.51	80.17	18.14	5.29	81.33	35.72	16.99	79.90	34.54	13.39	82.23	35.02	12.51	80.11
LwF (std)	±1.24	±1.15	±2.46	±0.66	±6.95	±7.53	±3.54	±5.19	±5.42	±2.29	±5.47	±4.37	±0.70	±3.58	±3.92
EWC	15.56	3.61	80.07	16.05	6.13	80.49	36.66	17.07	79.82	35.00	14.46	81.12	35.16	12.57	80.11
EWC (std)	±1.36	±1.45	±2.43	±2.05	±6.67	±7.13	±4.20	±5.07	±5.32	±2.59	±6.46	±5.58	±0.80	±3.74	±4.14
MAS	16.04	6.08	77.60	32.89	27.88	58.80	49.00	21.31	75.58	34.83	12.54	83.04	35.23	12.35	80.33
MAS (std)	±1.32	±2.00	±1.45	±5.76	±7.97	±7.27	±7.16	±5.98	±6.04	±1.61	±5.26	±4.14	±1.96	±3.78	±4.18
TWP	17.01	5.65	77.47	34.35	18.63	66.99	40.16	16.45	80.43	34.47	13.29	82.52	34.38	12.43	79.74
TWP (std)	±2.02	±1.84	±2.65	±9.25	±6.58	±6.22	±4.58	±3.35	±3.25	±1.62	±5.56	±4.14	±0.98	±3.76	±3.89
GFCIL	50.11	36.17	44.96	54.56	35.59	47.19	69.42	59.58	36.66	85.15	76.36	16.29	72.15	60.15	30.81
GFCIL (std)	±2.03	±2.73	±5.70	±5.27	±5.52	±4.82	±7.59	±6.63	±7.38	±3.33	±3.07	±5.01	±4.31	±5.30	±5.62
Geometer	47.55	30.15	50.28	39.19	22.45	47.78	77.45	57.80	38.82	77.68	62.16	28.61	65.79	44.97	44.03
Geometer (std)	±1.86	±3.08	±4.66	±3.59	±5.67	±5.59	±4.64	±6.66	±7.13	±3.36	±5.03	±5.71	±4.64	±4.31	±4.16
Mecoin	42.52	27.58	45.35	36.86	23.16	47.17	68.45	50.42	45.34	67.53	51.24	36.27	57.77	41.90	41.89
Mecoin (std)	±2.70	±1.47	±5.33	±2.07	±0.68	±3.82	±6.81	±4.79	±4.22	±1.66	±4.30	±4.11	±6.77	±8.31	±11.06
TAP	57.84	42.88	40.54	57.26	41.46	43.60	78.52	66.62	29.83	83.04	72.60	22.31	75.77	62.69	29.18
TAP (std)	±2.29	±1.53	±3.17	±5.39	±1.69	±3.69	±3.39	±3.94	±3.51	±5.00	±4.58	±3.66	±3.35	±5.49	±6.54
MAGIC	**62.92**	**52.29**	**27.91**	**62.07**	**47.82**	**36.36**	**93.57**	**90.58**	**6.40**	**85.44**	**79.09**	**11.65**	**78.44**	**71.74**	**17.00**
MAGIC (std)	**±0.79**	**±0.95**	**±3.81**	**±4.65**	**±3.00**	**±3.64**	**±0.71**	**±0.71**	**±1.13**	**±2.80**	**±2.58**	**±4.65**	**±2.98**	**±4.83**	**±6.10**

Table 8: Novel-class results (%) under the 1-shot setting. The mean and standard deviation over five runs are reported in consecutive rows. The best result in each column is shown in bold. MAGIC Backbone: SGC.

	CoraFull		ogbn-arxiv		Coauthor CS		Amazon Computers		WikiCS	
Method	AN	FN	AN	FN	AN	FN	AN	FN	AN	FN
LwF	16.22	3.27	11.26	6.84	43.70	22.85	47.56	33.32	46.27	23.89
LwF (std)	±2.05	±1.27	±4.13	±8.03	±2.48	±9.78	±10.84	±13.99	±4.14	±9.25
EWC	16.93	3.94	13.75	10.75	43.77	23.17	46.60	33.48	47.07	24.47
EWC (std)	±1.67	±2.26	±3.60	±6.83	±3.45	±9.73	±11.08	±13.98	±3.77	±9.73
MAS	20.32	8.79	7.99	8.17	58.28	30.91	59.58	32.25	50.77	26.07
MAS (std)	±2.48	±3.14	±1.22	±6.38	±8.85	±11.04	±7.87	±13.69	±7.63	±8.19
TWP	17.26	5.01	21.31	11.70	41.52	19.97	46.99	31.85	48.26	23.79
TWP (std)	±1.54	±2.14	±13.05	±6.12	±2.17	±8.52	±5.59	±14.49	±6.79	±9.88
GFCIL	5.48	5.89	21.70	15.42	12.38	10.35	37.60	35.24	17.03	17.75
GFCIL (std)	±0.89	±1.13	±8.01	±3.60	±1.78	±3.97	±16.67	±14.95	±6.20	±6.45
Geometer	8.99	3.94	15.25	5.31	39.97	27.65	30.32	27.40	26.98	17.21
Geometer (std)	±3.11	±1.20	±4.82	±5.35	±5.61	±10.96	±9.27	±14.82	±14.57	±13.07
Mecoin	14.04	11.21	14.76	12.53	32.12	25.31	46.74	34.08	23.03	19.11
Mecoin (std)	±3.01	±0.79	±6.02	±3.44	±6.48	±4.64	±7.73	±9.27	±9.10	±5.70
TAP	27.86	23.15	**35.16**	**27.92**	62.02	59.71	63.99	55.64	55.50	46.74
TAP (std)	±2.38	±3.36	**±6.53**	**±7.36**	±7.71	±6.98	±14.66	±10.45	±4.95	±3.92
MAGIC	**30.19**	**28.35**	31.54	24.23	**82.60**	**83.85**	**70.03**	**61.21**	**58.80**	**48.53**
MAGIC (std)	**±2.70**	**±2.47**	±7.00	±4.73	**±6.57**	**±4.66**	**±17.77**	**±16.90**	**±10.32**	**±8.95**

indicating that MAGIC effectively combines sparse labeled evidence with transferable topological relations. Its consistently strong FN further shows that the gains are retained throughout the incremental process rather than being limited to the sessions in which the classes first appear. Together with the overall results, these findings demonstrate a favorable balance between novel-class plasticity and continual retention.

B.7 Hyperparameter Sensitivity

We study the sensitivity of MAGIC on CoraFull under the 5-shot setting. The evaluated hyperparameters include the MRF relation strength η, pseudo-label mass ρ, distillation mass κ, number of prior bins B, distillation temperature T_D, unary strength τ_u, and maximum MRF subgraph hop. We vary one parameter at a time while keeping all other parameters at their default values. All curves report results averaged across five runs.

Table 9: Novel-class results (%) under the 3-shot setting. The mean and standard deviation over five runs are reported in consecutive rows. The best result in each column is shown in bold. MAGIC Backbone: SGC.

	CoraFull		ogbn-arxiv		Coauthor CS		Amazon Computers		WikiCS	
Method	AN	FN	AN	FN	AN	FN	AN	FN	AN	FN
LwF	22.19	4.99	15.97	6.85	45.59	23.23	57.41	33.81	49.51	26.29
LwF (std)	±1.20	±1.35	±3.74	±8.27	±3.44	±9.26	±12.08	±14.12	±9.38	±6.89
EWC	22.33	5.10	17.01	7.89	45.86	23.22	58.02	33.94	50.15	26.40
EWC (std)	±1.46	±1.80	±3.79	±7.96	±3.43	±9.13	±12.46	±14.22	±9.32	±7.20
MAS	22.38	8.59	9.94	8.05	57.42	28.54	59.77	31.85	49.37	26.04
MAS (std)	±1.17	±2.23	±1.65	±5.87	±3.55	±9.58	±8.10	±14.24	±8.24	±7.74
TWP	23.33	7.51	33.94	17.27	46.03	21.53	57.56	33.54	47.12	26.22
TWP (std)	±2.21	±1.74	±12.14	±7.66	±2.99	±4.42	±8.14	±14.30	±10.40	±7.68
GFCIL	16.44	16.74	35.22	22.86	43.45	48.13	66.11	57.10	27.87	30.40
GFCIL (std)	±1.07	±0.78	±6.80	±2.47	±5.96	±5.19	±12.96	±9.68	±5.71	±4.06
Geometer	15.02	10.01	22.15	12.02	63.85	45.94	43.84	24.95	31.92	15.38
Geometer (std)	±2.07	±1.26	±5.60	±7.56	±3.57	±5.96	±14.42	±11.90	±6.13	±4.87
Mecoin	20.59	16.99	23.98	17.86	48.04	40.75	45.71	36.98	33.45	26.98
Mecoin (std)	±3.15	±1.33	±6.32	±2.72	±6.94	±4.61	±5.26	±8.29	±4.96	±4.15
TAP	38.63	32.19	41.39	31.92	67.97	64.21	71.15	64.11	61.19	53.76
TAP (std)	±3.21	±2.15	±7.81	±6.14	±3.05	±3.54	±8.55	±5.45	±9.23	±7.55
MAGIC	**50.20**	**45.27**	**45.73**	**34.59**	**90.89**	**89.59**	**78.17**	**69.42**	**64.45**	**63.13**
MAGIC (std)	**±1.75**	**±1.45**	**±6.30**	**±4.65**	**±1.79**	**±0.35**	**±9.91**	**±9.92**	**±9.77**	**±8.00**

Table 10: Novel-class results (%) under the 5-shot setting. The mean and standard deviation over five runs are reported in consecutive rows. The best result in each column is shown in bold. MAGIC Backbone: SGC.

	CoraFull		ogbn-arxiv		Coauthor CS		Amazon Computers		WikiCS	
Method	AN	FN	AN	FN	AN	FN	AN	FN	AN	FN
LwF	23.86	5.41	15.06	6.01	46.25	23.51	60.18	34.25	51.46	27.22
LwF (std)	±0.88	±1.72	±3.72	±6.89	±4.24	±9.79	±8.12	±14.62	±7.31	±8.14
EWC	23.88	5.47	16.26	7.57	46.48	23.71	59.76	34.16	51.65	27.16
EWC (std)	±0.95	±1.65	±3.72	±8.06	±4.32	±9.88	±7.98	±14.46	±7.30	±8.00
MAS	23.51	8.60	10.06	7.04	55.78	28.19	60.05	33.65	49.31	26.31
MAS (std)	±1.31	±2.18	±3.04	±3.87	±5.72	±9.97	±8.17	±13.90	±7.94	±8.63
TWP	24.32	7.07	32.41	16.16	46.63	21.38	59.14	33.78	51.71	27.97
TWP (std)	±1.68	±2.05	±9.86	±5.62	±4.06	±6.76	±7.71	±14.36	±7.01	±7.65
GFCIL	25.54	24.70	41.07	30.64	59.95	63.14	72.41	63.72	45.96	47.45
GFCIL (std)	±0.47	±1.00	±5.13	±3.42	±4.24	±5.04	±7.81	±6.07	±5.43	±2.99
Geometer	20.86	15.18	20.77	4.93	71.34	64.10	48.31	30.82	42.64	22.16
Geometer (std)	±2.09	±2.03	±4.98	±1.51	±5.50	±4.88	±8.58	±11.36	±11.09	±9.26
Mecoin	25.46	21.66	32.49	24.18	54.25	44.63	49.29	40.92	43.27	33.42
Mecoin (std)	±2.12	±1.48	±9.93	±4.83	±4.08	±5.25	±4.42	±11.47	±7.36	±4.29
TAP	42.96	36.43	42.77	33.78	69.84	65.28	72.41	66.94	62.26	56.04
TAP (std)	±1.61	±2.15	±7.90	±6.27	±2.57	±3.96	±8.03	±5.88	±9.19	±8.09
MAGIC	**56.56**	**51.81**	**48.71**	**37.49**	**92.02**	**90.87**	**82.27**	**73.67**	**69.89**	**68.81**
MAGIC (std)	**±2.27**	**±2.19**	**±5.10**	**±5.53**	**±1.70**	**±0.70**	**±6.49**	**±10.24**	**±8.00**	**±5.96**

Figure 6 shows that MAGIC is generally robust over broad hyperparameter ranges. Increasing the relation strength η moderately improves overall and final accuracy, while the novel-class metrics remain largely stable. This indicates that stronger relational evidence is beneficial without causing unstable topology-driven propagation. The pseudo-label mass ρ has a more pronounced effect: increasing it substantially improves all metrics, especially novel-class accuracy, with the gains gradually saturating at larger values. This confirms the importance of MRF-derived auxiliary supervision while showing that its influence is well controlled by class-balanced mass allocation.

The distillation mass κ exhibits a clear retention–plasticity trade-off. Moderate values preserve overall performance, whereas excessive distillation increasingly suppresses novel-class learning. The final overall accuracy remains relatively stable over moderate settings but decreases when historical supervision becomes dominant. For the distillation temperature T_D, lower temperatures slightly improve overall and final accuracy by providing sharper teacher targets, while higher temperatures produce a small increase in mean novel-class accuracy. Final novel-class performance is comparatively insensitive to this parameter.

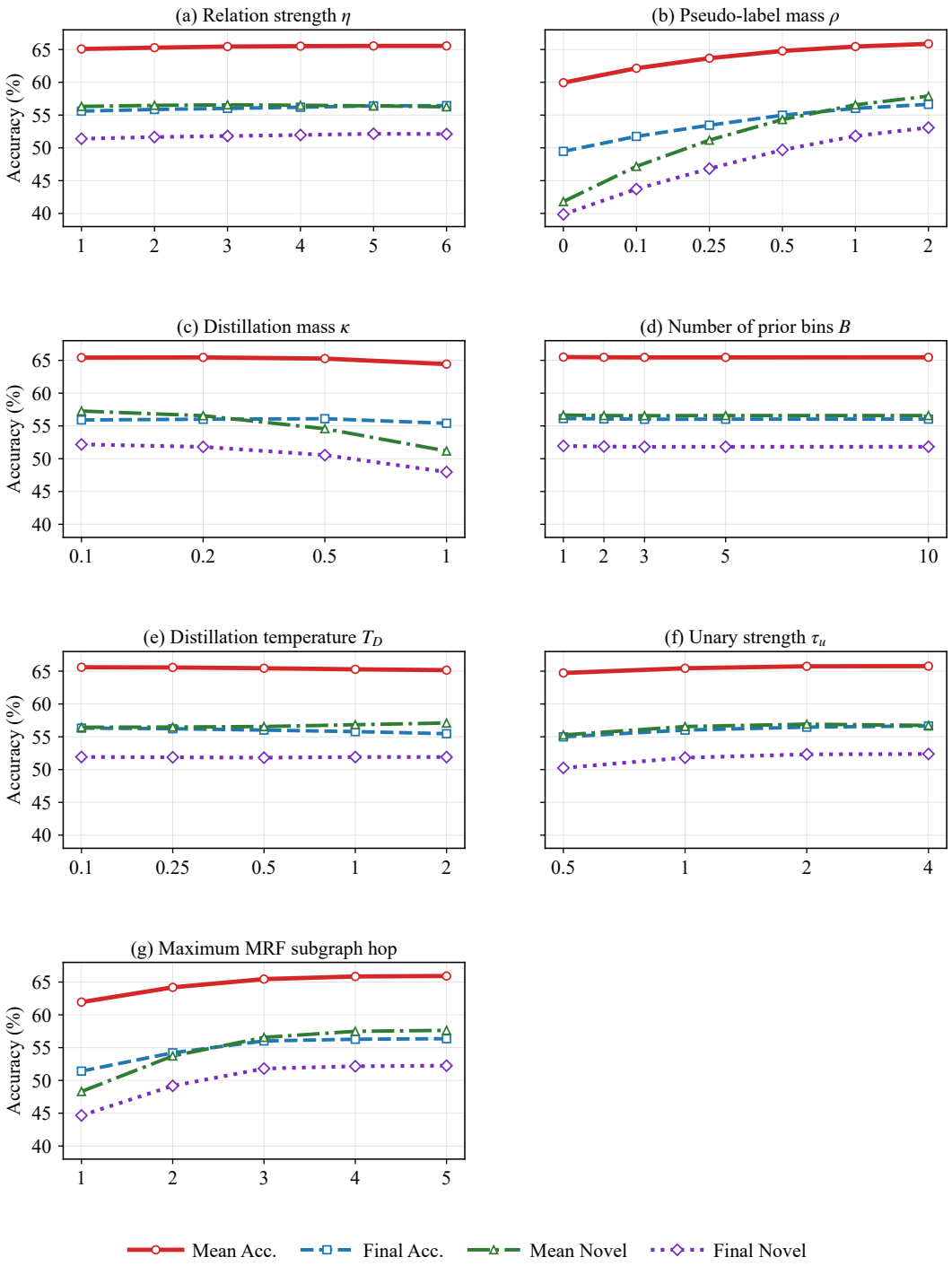


Figure 6: Hyperparameter sensitivity of MAGIC on CoraFull under the 5-shot setting. Each subplot varies one hyperparameter while fixing all others to their default values. MAGIC Backbone: SGC

MAGIC is nearly insensitive to the number of prior bins B, suggesting that the class-balanced relation prior remains reliable under both coarse and fine similarity discretizations. Increasing the unary strength τ_u improves performance up to a moderate level, after which the results largely saturate, demonstrating robustness to the relative weighting of classifier evidence and pairwise relations. Finally, expanding the MRF subgraph from one to three hops yields substantial improvements, while

Table 11: Hyperparameter configurations used in the main experiments. The same configurations are used for the 1-, 3-, and 5-shot settings.

Dataset	η	ρ	κ	B	T_D	τ_u	k
CoraFull	3.0	1.0	0.2	3	0.50	1.0	3
ogbn-arxiv	6.0	1.0	0.5	10	0.25	1.0	3
Coauthor CS	1.0	1.0	0.5	10	0.25	1.0	3
Amazon Computers	1.0	1.0	0.5	10	0.25	1.0	3
WikiCS	3.0	1.0	0.5	10	0.25	1.0	3

the gains beyond three hops are small. Overall, the default configuration provides a balanced compromise between relational propagation, pseudo-label supervision, historical retention, and novel-class adaptation.

C LIMITATIONS

MAGIC has two main limitations. First, it does not explicitly determine whether the topology prior learned in the base session will have a positive effect in each novel session. When the relation patterns of novel classes differ substantially from those observed in the base session, injecting the transferred prior may provide limited or even misleading supervision. Future work could develop a session-level prior-validity estimator that adaptively measures transferability and adjusts the strength of prior injection.

Second, MRF inference requires access to unlabeled nodes and edges surrounding the support set. MAGIC is therefore most applicable to transductive or dynamically visible GFSCIL settings. Its topology-based prior injection is more difficult to apply in strictly inductive, privacy-constrained, or topologically incomplete scenarios, where the relevant neighborhood structure is unavailable or only partially observed.

D REPRODUCIBILITY DETAILS

All experiments were conducted on a 64-bit Windows system (build 10.0.26200) equipped with an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of memory. The software environment used Python 3.10.20, PyTorch 2.5.1 compiled with CUDA 12.1, PyTorch Geometric 2.8.0, and NumPy 2.2.6. The NVIDIA driver version was 596.08. CUDA was enabled throughout the experiments. Each result was averaged over five runs with random seeds $\{0, 1, 2, 3, 4\}$, and we report the mean and standard deviation across these runs.

The random seed controls class and support splits, prior-node sampling, and the fixed random feature projection. For every dataset, the same splits are used across the 1-, 3-, and 5-shot settings. Hyperparameters of MAGIC for each dataset in the main experiments are provided in Table 11.