

# LoRETTA: A Low Resource Framework To Poison Continuous Time Dynamic Graphs

Himanshu Pal\*, Venkata Sai Pranav Bachina\*, Ankit Gangwal, Charu Sharma

International Institute of Information Technology, Hyderabad

## Abstract

Temporal Graph Neural Networks (TGNNs) are increasingly used in high-stakes domains, such as financial forecasting, recommendation systems, and fraud detection. However, their susceptibility to poisoning attacks poses a critical security risk. We introduce **LoRETTA** (**L**ow **R**esource **T**wo-phase **E**nterprise **T**emporal **A**ttack), a novel adversarial framework on Continuous-Time Dynamic Graphs, which degrades TGNN performance by an average of **29.47%** across 4 widely benchmark datasets and 4 State-of-the-Art (SotA) models.

LoRETTA operates through a two-stage approach: (1) sparsify the graph by removing high-impact edges using any of the 16 tested temporal importance metrics, (2) strategically replace removed edges with adversarial negatives via LoRETTA’s novel degree-preserving negative sampling algorithm. Our plug-and-play design eliminates the need for expensive surrogate models while adhering to realistic unnoticeability constraints. LoRETTA degrades performance by upto **42.0%** on MOOC, **31.5%** on Wikipedia, **28.8%** on UCI, and **15.6%** on Enron. LoRETTA outperforms 11 attack baselines, remains undetectable to 4 leading anomaly detection systems, and is robust to 4 SotA adversarial defense training methods, establishing its effectiveness, unnoticeability, and robustness.

**Code** — <https://github.com/ansh997/LoReTTA>

## 1 Introduction

Temporal Graph Neural Networks (TGNNs) are increasingly deployed in real-world systems such as social media analysis (Deng, Rangwala, and Ning 2019; Fan et al. 2019), transportation modeling (Jiang and Luo 2022; Yu, Yin, and Zhu 2017), recommendation engines (Gao et al. 2022; Wu et al. 2022), outbreak tracking (Cencetti et al. 2021; So et al. 2020), and knowledge graph reasoning (Cai et al. 2022). These applications often rely on modeling evolving relationships over time using dynamic graphs, making the robustness of TGNNs a crucial concern.

Dynamic graphs can be broadly categorized as Discrete-Time Dynamic Graphs (DTDGs) and CTDGs. DTDGs represent dynamic graphs as a sequence of static snapshots, while CTDGs model them as a continuous stream of timestamped interactions. CTDGs are considered more expressive

for modeling real-world systems, as they capture fine-grained temporal dependencies that DTDGs may miss (Zheng, Wei, and Liu 2023; Ennadir et al. 2024).

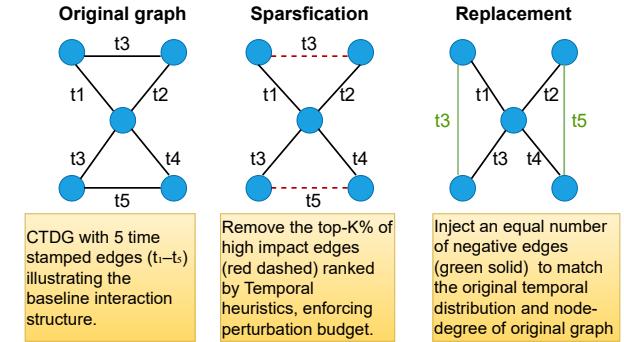


Figure 1: Overview of the LoRETTA attack framework.

Recent work (Lee, Lee, and Shin 2024) has demonstrated that TGNNs are susceptible to adversarial poisoning attacks, i.e small but carefully crafted perturbations to the training graph that can significantly degrade performance. This threat becomes especially critical in high-stakes domains such as fraud detection or traffic forecasting, where reliability is paramount.

While adversarial attacks have been extensively studied for static graphs (Zügner, Akbarnejad, and Günnemann 2018; Bojchevski and Günnemann 2019; Dai et al. 2018; Ma, Ding, and Mei 2020a; Zügner and Günnemann 2024), their temporal counterparts (CTDGs) are less explored. The challenges are non-trivial: CTDGs evolve over time, requiring the attacker to perturb edges with precise temporal coordination. TGNNs further complicate this by maintaining memory states that propagate historical information and may dilute isolated attacks. Moreover, the attacker must decide *when* and *where* to insert perturbations to maximize the impact, all the while adhering to unnoticeability constraints (cf §4).

The SotA work, i.e., Spear and Shield (Lee, Lee, and Shin 2024), introduces an adversarial poisoning framework for TGNNs, but it carries significant practical limitations. Specifically, it relies on training a surrogate model to estimate gradients, an approach that is **computationally expensive** for

\*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

**large-scale datasets.** Furthermore, it assumes the attacker has complete access to the entire dataset (train, validation, and test) and the ability to make arbitrary changes. In real-world scenarios, such assumptions rarely hold in the context of an attacker. Attackers often only have access to the training portion of the graph and must operate under realistic, limited knowledge and manipulation constraints.

To address these limitations, we propose a novel adversarial poisoning framework for TGNNs, that we call **LORETTA** (**L**ow **R**esource **T**wo-phase **T**emporal **A**ttack). LORETTA does not rely on surrogate models, and it assumes access to only the training portion of the data. It adheres to 4 practical unnoticeability constraints (cf. §4) that preserve the structural and temporal plausibility of the poisoned graph. At its core, LORETTA computes scores for edges (using any of our 16 tested heuristics) across timestamps to quantify evolving influence, and removes edges with high temporal influence; thereby maximizing disruption to the TGNN’s representation learning. These edges are then replaced with strategically selected adversarial negatives through LORETTA’s novel negative sampling algorithm designed to preserve the original graph’s edge density and local structural cues.

Our main contributions are summarized below:

1. We propose LORETTA, a novel adversarial poisoning framework that operates by removing high-influence edges and replacing them with adversarial negatives using LORETTA’s novel negative sampling algorithm.
2. We enforce 4 unnoticeability constraints (cf. §4) that includes graph sparsity, temporal plausibility, node activity windows, and degree preservation to ensure attack stealth.
3. We demonstrate LORETTA (a) outperforms 11 baselines (cf. §7.1) with 29.47% average degradation across 4 datasets and 4 SotA TGNN models; (b) while evading 4 SotA anomaly detectors (cf. §7.3) and (c) resisting 4 adversarial defense methods (cf. §7.2).

## 2 Related Work

### 2.1 Temporal Graph Neural Networks (TGNNs)

Dynamic graphs are commonly categorized into two settings: DTDGs and CTDGs. DTDG approaches discretize a temporal graph into sequential snapshots and apply static graph learning to each frame (Pareja et al. 2020; Sankar et al. 2020). However, these methods neglect fine-grained temporal continuity, leading to less faithful modeling of real-world dynamic systems. In contrast, CTDG-based approaches, particularly TGNNs (Trivedi et al. 2019; Rossi et al. 2020; Xu et al. 2020; Ma et al. 2020), model interactions as continuous temporal streams and capture temporal dependencies directly. These models often incorporate memory networks (Ma et al. 2020; Rossi et al. 2020), sequential encoders (Cong et al. 2023), and temporal message-passing schemes, allowing more expressive modeling of evolving systems. TGNNs have shown success in modeling complex phenomena such as user-item interactions, social dynamics, and temporal knowledge graphs (Ennadir et al. 2024). In this work, we focus on CTDGs due to their expressive power and alignment with real-world interaction patterns.

### 2.2 Adversarial Attacks on Graphs

Adversarial attacks fall into two broad categories:

*Evasion Attacks:* Manipulates inputs at inference time to mislead the model after training.

*Poisoning attacks:* Injects perturbations *during training* to corrupt the learned representations. **LORETTA is a poisoning attack**, aiming to degrade overall model performance by manipulating the training graph.

**Poisoning static graphs vs. temporal graphs:** A substantial body of literature explores poisoning attacks on static graphs (Bojchevski and Günnemann 2019; Li et al. 2020), where edge or node feature perturbations can significantly impact downstream tasks like node classification or link prediction. However, directly applying these methods to CTDGs is ineffective due to the evolving nature of temporal graphs. Specifically, past perturbations quickly lose influence as new edges arrive, and future edges are not observable at attack time, preventing attackers from crafting precise interventions. This dynamic nature limits the transferability of static attack strategies to temporal domains.

**Poisoning DTDGs vs CTDGs:** Chen et al. (2021); Sharma et al. (2023) extend poisoning strategies to DTDGs by targeting discrete graph snapshots. While effective in short-term settings, these methods are inherently limited in scope. Perturbations in DTDGs only influence the snapshot in which they occur and lack temporal propagation, making them ill-suited for CTDGs where edge interactions carry real-valued timestamps and temporal continuity is essential.

**Existing Literature on Poisoning CTDGs:** T-SPEAR (Lee, Lee, and Shin 2024) is, to our knowledge, the first dedicated poisoning attack on CTDGs. It trains a surrogate model to select edges that, when inserted, disrupt downstream predictions while adhering to 4 unnoticeability constraints (similar but weaker than ours). They also present an adversarial defense strategy against poisoning attacks called T-shield. While effective, T-SPEAR has critical limitations: (1) Assumes full access to the dataset, which is not true in real-world scenarios, as adversary often has access to just training dataset. (2) Requires significant compute to train and optimize a surrogate model. (3) Introduces edges but does not consider deleting critical ones, limiting perturbation scope. (4) The effectiveness of the attack depends on the accuracy of the surrogate model used.

## 3 Preliminaries

**Dynamic graph setting:** We consider a CTDG defined as  $G = (V, E)$ , where  $V$  is the set of  $n$  nodes, and  $E$  is the set of  $m$  timestamped edges. Each edge is a tuple  $(u, v, t)$ , where  $u, v \in V$  denote the interacting nodes and  $t \in \mathbb{R}$  is the time of interaction. Multiple edges can exist between the same node pair at different times, capturing repeated interactions over time.

**Temporal PageRank:** Temporal PageRank (TPR) extends classical PageRank to temporal graphs by replacing static walks with time-respecting walks (Rozenshtein and Gionis 2016). Let  $Z^T(v, u \mid t)$  denote the set of temporal walks (cf

Appendix A.12) from node  $v$  to  $u$  that occur strictly before time  $t$ . The probability of a walk  $z$  is defined as:

$$\Pr[z \in Z^T(v, u | t)] = \frac{c(z | t)}{\sum_{z' \in Z^T(v, x | t), x \in V, |z'|=|z|} c(z' | t)},$$

where  $c(z | t)$  is the decay-weighted count of  $z$ :

$$c(z | t) = (1-\beta) \prod_{\substack{((u_{i-1}, u_i, t_i), \\ (u_i, u_{i+1}, t_{i+1})) \in z}} \beta^{|\{(u_i, y, t') | t' \in [t_i, t_{i+1}], y \in V\}|}.$$

Transitions are penalized exponentially by the number of intervening interactions to model temporal decay. The final TPR score of node  $u$  at time  $t$  is:

$$r(u, t) = \sum_{v \in V} \sum_{k=0}^t (1-\alpha) \alpha^k \sum_{\substack{z \in Z_T(v, u | t) \\ |z|=k}} \Pr[z | t],$$

where  $\alpha$  is the jump probability. This formulation biases the walk toward shorter, temporally coherent paths and naturally captures influence drift in evolving graphs.

## 4 Problem Formulation

**Attacker’s objective.** Given a clean CTDG  $G$ , the attacker seeks to degrade the performance of a Temporal Graph Neural Network (TGNN) on dynamic link prediction by injecting adversarial edges and/or removing crucial ones. The perturbed graph is denoted as:

$$\tilde{G} = (V, (E \setminus E') \cup \tilde{E}),$$

where  $E' \subset E$  are the removed edges,  $\tilde{E}$  are the adversarial (inserted) edges, and the total number of modifications is bounded by a fixed budget  $\Delta$ . This is a **poisoning attack**, the graph is modified before training, and the TGNN is trained on the corrupted  $\tilde{G}$ .

**Attacker’s knowledge.** We assume a strict black-box setting: the attacker has no knowledge of the TGNN architecture, loss function, or gradients. However, the attacker can observe the training portion of the dataset, a common assumption in graph poisoning (Lee, Lee, and Shin 2024; Ma, Ding, and Mei 2020b; Zigner, Akbarnejad, and Günnemann 2018). Unlike previous work like T-spear (Lee, Lee, and Shin 2024), we do not assume access to validation/test splits.

**Unnoticeability Constraints** To ensure the attack remains stealthy and realistic, we adopt and strengthen the four unnoticeability constraints introduced by T-spear:

- **(C1) Perturbation Budget:** The total number of modified edges (inserted + removed) is bounded by  $\Delta = [p \cdot |E|]$ , where  $p$  is the perturbation rate.
- **(C2) Temporal Feasibility:** Timestamps of inserted edges are drawn from the same distribution as original timestamps, i.e.,  $\tilde{t} \sim P_t(E)$ .

- **(C3) Node Activity Window:** Inserted edges may only connect nodes that have been active within a temporal window  $W$  around the chosen timestamp  $\tilde{t}$ . That is, both endpoints of  $\tilde{e} = (u, v, \tilde{t})$  must appear in  $V_{i,W} = \{u_j, v_j | j \in [i-W+1, i]\}$ .

- **(C4) Degree Preservation:** We impose a stronger variant of node-level stealth by matching the degree distribution of each node before and after perturbation. This prevents attackers from noticeably increasing or decreasing a node’s interaction footprint.

## 5 Methodology

We propose a two-stage adversarial attack framework, **Sparsification** followed by **Replacement**, designed to degrade TGNNs trained on CTDGs. In this section, we explain LORETTA in detail.

### 5.1 Step 1: Sparsification

We perform the Sparsification of a CTDG in 16 different ways. We classify them broadly into broad types:

**Edge Sparsification Strategy:** For each timestamp  $t_i$ , we construct a static aggregated graph  $G^{(i)} = (V, E^{(i)})$  where  $E^{(i)} = \{(u, v, t) \in E : t \leq t_i\}$ . We define a general temporal importance function  $H : V \times V \times \mathbb{R} \rightarrow \mathbb{R}$  that computes the importance score for any edge  $(u, v, t_j)$  as:

$$H(u, v, t_j) = f((u, v), G^{(j)}), \quad (1)$$

where  $f$  represents any graph-theoretic heuristic applied to edge  $(u, v)$  within the context of static aggregation  $G^{(j)}$ . Edges are ranked by  $H(u, v, t_j)$  in descending order, and the top- $\Delta$  highest scoring edges are selected for removal to achieve the desired  $p$ . We consider 4 heuristic strategies and also a Random edge removal strategy (cf Appendix A.1).

**Timestamp Sparsification Strategy:** We identify critical timestamps by measuring temporal drift in node importance through generalized distance metrics. Let  $r^{(t_i)} \in \mathbb{R}^{|V|}$  denote the TPR vector at timestamp  $t_i$ . We define temporal drift at timestamp  $t_i$  as:

$$\delta^{(t_i)} = d(r^{(t_i)}, r^{(t_{i-1})}), \quad (2)$$

where  $d(\cdot, \cdot)$  represents a generalized distance function measuring importance shift between consecutive timestamps. High  $\delta^{(t_i)}$  values indicate volatile periods where minimal perturbations significantly impact learned temporal dynamics. We rank all timestamps by their drift values and select the top timestamps with highest  $\delta^{(t_i)}$  scores. If the final timestamp exceeds  $\Delta$ , we select exactly the remaining number of edges at that timestamp to meet the budget. Similar to edge sparsification, where heuristic scores can be replaced with various graph-based metrics, temporal drift computation supports multiple distance metrics (eg.  $\ell_2$ -norms, cosine distance, Jaccard distance etc). We consider with 11 different distance metrics (cf Appendix A.1).

## 5.2 Step 2: Adversarial Negative Sampling

After removing critical edges in Step 1, we insert feasible negative edges to maintain graph density and temporal dynamics. This is designed to respect unnoticeability constraints (C2–C4) without requiring access to model gradients or internals. Our negative sampling algorithm operates as follows: for each removed edge, we insert a new edge that (i) mimics natural temporal patterns, (ii) connects temporally active nodes, and (iii) preserves the degree distribution.

**C2:** We first sample candidate timestamps from the empirical distribution of existing edge times using kernel density estimation (KDE). This ensures inserted edges follow the same temporal rhythm as genuine interactions, satisfying temporal stealth (**C2**).

**C3:** Given a sampled timestamp, we construct a candidate node pool by selecting nodes active within a local time window  $W$  around that timestamp. This enforces the activity constraint (**C3**), ensuring inserted edges do not connect inactive or dormant nodes, which would appear suspicious under standard monitoring tools. We adopt the same  $W$  used by T-spear (Lee, Lee, and Shin 2024). For bipartite datasets, endpoints are sampled from disjoint node sets. Additionally, we enforce that node pairs  $(u, v)$  are new to the graph—no prior interaction exists in either direction—preventing semantic leakage and contradictions with observed history.

**C4:** To maintain degree consistency (**C4**), we track deletions and insertions per node and update candidate pools accordingly. New edges are selected to keep each node’s in-degree and out-degree statistically unchanged, guarding against structural anomalies such as hub overloading or connectivity spikes.

Our constraint-aware algorithm iteratively samples valid edge–timestamp pairs that satisfy all stealth constraints. When no valid candidates remain due to constraint overlap or capacity limits, we trigger a recovery step that reinitializes the KDE over timestamps and samples a fresh set of candidate times. This resampling enables the algorithm to explore new feasible timestamps without relaxing any constraints, ensuring the perturbation budget is met while preserving unnoticeability.

These steps form a principled, constraint-driven negative sampling routine. Unlike learned adversarial strategies requiring surrogate models, our method generates structurally and temporally plausible perturbations without supervision, yielding effective attacks in low-resource settings. Alternative approaches like random sampling (selecting edges arbitrarily out of temporal order) or Havel-Hakimi (Kleitman and Wang 1973) are insufficient—random sampling lacks strategic impact, while Havel-Hakimi preserves degrees without addressing temporal dynamics or unnoticeability requirements. We provide the pseudocode for LORETTA’s negative sampling algorithm in Appendix A.14.

## 5.3 Complexity Analysis

We analyze the time and space complexity of each component in LORETTA.

**Time complexity of Temporal PageRank** *The time complexity of the TPR algorithm is  $O(|E| + |V|)$ , where  $|E|$  is*

*the number of edges in the input graph and  $|V|$  is the number of vertices.*

**Space Complexity of Temporal PageRank** *Let  $|V|$  be the number of distinct nodes and  $|T|$  the number of distinct timestamps in the input sequence. The combined data structures of TPR and TER occupy*

$$O(|V| |T|) \text{ memory.}$$

**Time Complexity of TimeStamp Selector** *The time complexity of our method is*

$$O(|V|(\log |V| + d_{\max} k \log k) + |E| \log |E|),$$

*where  $|V|$  the number of vertices  $|E|$  is the number of edges,  $d_{\max}$  is the maximum degree of any vertex, and  $k$  is the average number of timestamps per edge.*

**Space complexity of Timestamp Selector** *Let  $W$  be the length of the time window used when gathering candidate edges. Algorithm Timestamp Selector uses*

$$O(|V| + |E| W) \text{ memory.}$$

Detailed proofs and empirical analysis are provided in Appendix A.5, where we also demonstrate that LORETTA achieves up to 10× speedup over T-spear.

## 6 Experiment Setup

We provide details about datasets and models along with metrics and baselines used in experiments to show the effectiveness of our method (cf. Appendix A.9).

**Datasets:** We perform experiments on 4 benchmark real-world datasets: Wikipedia (Kumar, Zhang, and Leskovec 2019), MOOC (Feng, Tang, and Liu 2019), UCI (Panzarasa, Opsahl, and Carley 2009), and Enron (Shetty and Adibi 2004). For all datasets, we adopt a consistent chronological split of 70%-15%-15% for training-validation-test sets, following the methodology in (Rossi et al. 2020). Detailed descriptions of these datasets are provided in Appendix A.9.

**Models:** To evaluate the effectiveness of LORETTA, we conduct experiments on 4 SotA TGNNS: TGN (Rossi et al. 2020), JODIE (Kumar, Zhang, and Leskovec 2019), TGAT (Xu et al. 2020), and DySAT (Sankar et al. 2018).

**Metrics:** To ensure consistency and comparability, we utilize Mean Reciprocal Rank (MRR), which is a robust and a widely recognized evaluation metric for ranking tasks.

**Baselines:** We evaluate our method against 11 baselines: T-spear (SotA CTDG poisoning attack), 5 edge addition baselines, i.e., ADD, which were used in the T-spear paper, and 5 edge removal baselines, i.e., REM, extrapolated from static graph methods. Unlike the original T-spear work, which poisons the entire dataset, including the validation and test sets, we poison only the training set even for T-spear to ensure a fair comparison with our method, which accounts for any differences in reported T-spear performance numbers. Detailed explanation of baselines is in Appendix A.11.

## 7 Discussion

In this section, we evaluate performance of LORETTA against baselines (cf. §7.1), highlight its robustness to SotA defenses (cf. §7.2), anomaly detection algorithms (cf. §7.3) and present our ablation studies(cf. §7.4)

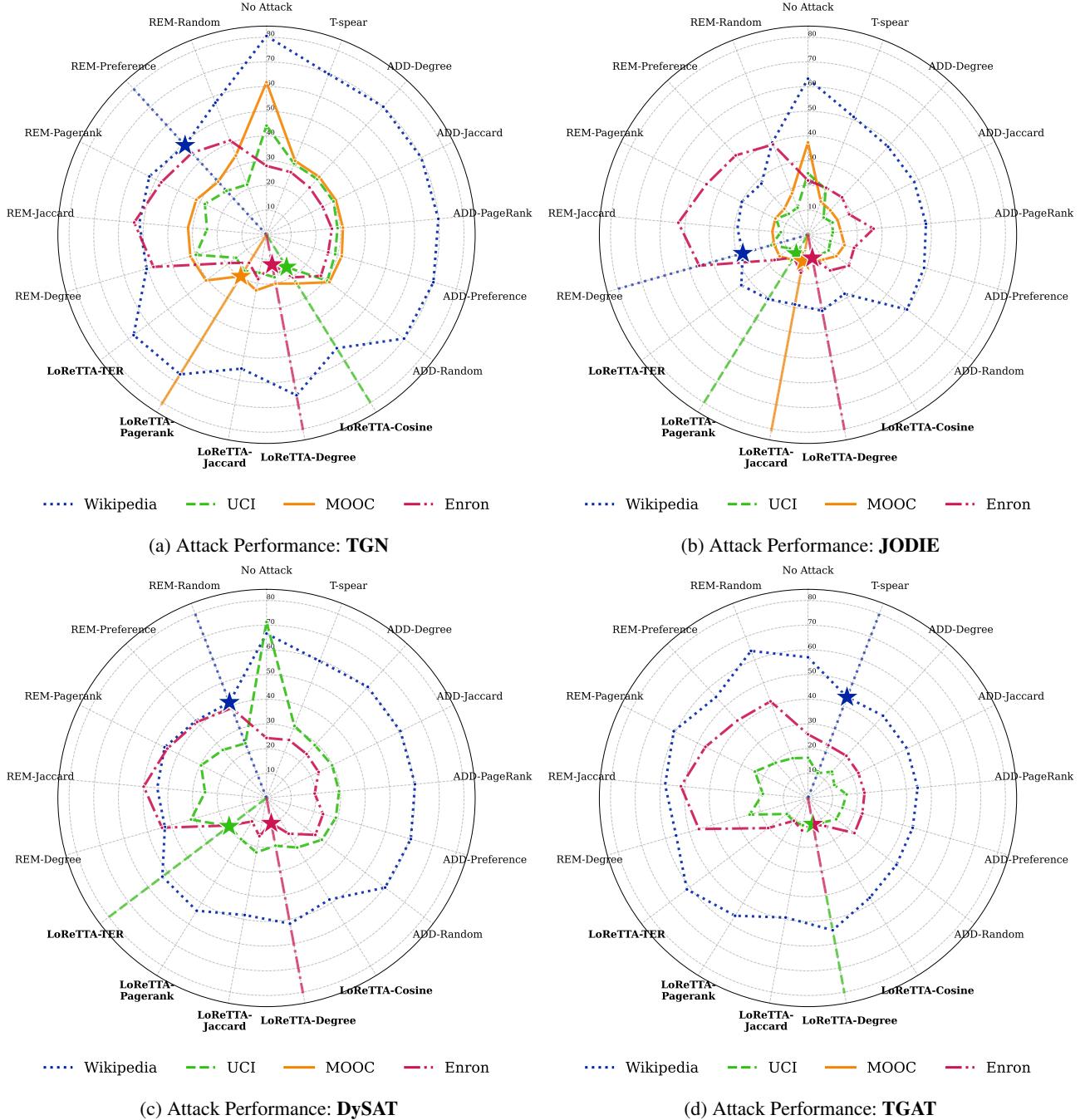


Figure 2: Performance of our attack on each model across datasets. Lower values indicate stronger attack impact. Clean baselines are shown under No-Attack. DySAT and TGAT fail on MOOC due to Out-Of-Memory error. Naive Jaccard is omitted for bipartite graphs (MOOC and UCI), where it is not applicable. Dashed radial lines highlight the most effective attack for each dataset-model pair. LORETTA-Degree is the best metric for both Enron and UCI on TGAT.

## 7.1 Attack Performance

We perform a comprehensive evaluation of our proposed attack method against 11 baseline approaches (cf Appendix A.11). Figure 2 presents the average performance of 5 runs of all victim models under various attack methods at a perturbation rate of 0.3. The results for MOOC on TGAT and

DySAT are not included due to out-of-memory errors caused by the size of the models.

LORETTA consistently beats all the baselines in the majority of cases, demonstrating superior attack effectiveness. The single exception occurs on the Wikipedia dataset, where REM baselines achieve comparable or slightly better perfor-

Attack	SVD		Cosine		T-shield-F		T-shield	
	Wikipedia	UCI	Wikipedia	UCI	Wikipedia	UCI	Wikipedia	UCI
Degree	40.96 $\pm$ 1.29	34.04 $\pm$ 1.78	27.57 $\pm$ 1.05	8.05 $\pm$ 0.37	10.45 $\pm$ 0.65	5.78 $\pm$ 0.62	10.30 $\pm$ 0.90	5.98 $\pm$ 0.32
PageRank	40.27 $\pm$ 0.88	34.44 $\pm$ 1.26	27.55 $\pm$ 1.40	7.72 $\pm$ 0.38	10.98 $\pm$ 0.74	5.86 $\pm$ 0.40	11.03 $\pm$ 0.80	5.62 $\pm$ 0.41
Cosine	46.01 $\pm$ 1.36	34.34 $\pm$ 1.26	17.87 $\pm$ 2.11	7.63 $\pm$ 0.90	10.32 $\pm$ 0.57	6.25 $\pm$ 0.41	9.62 $\pm$ 1.08	6.10 $\pm$ 0.79
Jaccard	48.94 $\pm$ 0.97	33.70 $\pm$ 1.69	19.33 $\pm$ 2.19	7.74 $\pm$ 0.57	9.08 $\pm$ 0.74	5.69 $\pm$ 0.52	9.61 $\pm$ 0.37	5.50 $\pm$ 0.27
TER	47.99 $\pm$ 1.39	34.46 $\pm$ 0.94	32.11 $\pm$ 0.97	7.40 $\pm$ 0.76	11.31 $\pm$ 1.07	6.20 $\pm$ 0.57	11.38 $\pm$ 1.05	5.96 $\pm$ 0.41
Clean	80.50 $\pm$ 0.50	44.2 $\pm$ 0.4	80.50 $\pm$ 0.50	44.2 $\pm$ 0.4	80.50 $\pm$ 0.50	44.2 $\pm$ 0.4	80.50 $\pm$ 0.50	44.2 $\pm$ 0.4

Table 1: Performance (%) MRR) of variants of LORETTA against 4 SotA defenses. The "Clean" row establishes the baseline performance of TGN without any perturbations or defenses applied.

mance. This anomaly stems from Wikipedia’s unique structural properties that make sparsification-only attacks particularly effective. Specifically, Wikipedia exhibits highly modular semantic clusters where users contribute predominantly to domain-specific areas. Sparsification attacks exploit this modularity by removing high-signal edges within semantic clusters, directly undermining the model’s ability to capture critical local coherence patterns. In contrast, our method’s negative sampling component introduces cross-domain edges that inadvertently provide implicit regularization, encouraging broader generalization across otherwise disjoint clusters and partially counteracting the sparsification damage. However, it should be noted that REM baselines achieve this performance by adhering only to the C1 constraint, thereby disrupting CTDG structure more extensively than our approach, which respects both C1 and C2 constraints. We provide a detailed analysis of this phenomenon in Appendix A.15.

Importantly, LORETTA demonstrates clear superiority across all other datasets, proving our method’s general effectiveness. Furthermore, LORETTA doesn’t just work for a specific sparsification heuristic. We consistently outperform many baselines for all 5 chosen sparsification strategies, showing the stability of our attack framework and its independence from specific sparsification techniques. The attack performance in tabular form is in Appendix A.2. We also assess dominance using a one-sided exact binomial sign test across baselines; see Appendix A.19 and Table 11.

## 7.2 Robustness against SotA Defenses

We evaluate the robustness of our proposed LORETTA against 4 SotA adversarial defense methods using TGN on Wikipedia and UCI datasets with  $p$  as 0.3, as presented in Table 1. The defenses include 2 static graph methods (SVD-based reconstruction and cosine similarity filtering) and 2 variations of Tshield (Lee, Lee, and Shin 2024) (Tshield-F and Tshield), which represent the current SotA CTDG defense methods.

Our results demonstrate that even after applying these defense mechanisms, victim models consistently fail to recover their original clean performance levels. Notably, we observe that Cosine similarity filtering, Tshield, and Tshield-F defenses cause additional performance degradation beyond that induced by LORETTA alone. This counterintuitive result occurs because these filtering-based approaches attempt to identify and remove adversarial edges from the dataset. However, when adversarial edges are incorrectly identified, true edges are inadvertently removed, resulting in further

performance deterioration. We experimentally validate this hypothesis in Appendix A.17, where we discovered that only approximately 30% of the filtered edges correspond to actual adversarial modifications. This finding not only validates the unnoticeability of LORETTA but also underscores its robustness against current defense strategies, highlighting the urgent need for more robust defense methods

## 7.3 Robustness against Anomaly Detection methods

Attack	Anomaly Detection Methods							
	MIDAS		F-FADE		AnoEdge-L		AnoEdge-G	
	P	R	P	R	P	R	P	R
<b>Wikipedia Dataset</b>								
Degree	0.35	0.52	1.00	0.28	0.47	0.45	0.39	0.55
PageRank	0.35	0.53	1.00	0.30	0.49	0.46	0.40	0.55
Cosine	0.26	0.70	1.00	0.21	0.50	0.38	0.46	0.74
Jaccard	0.28	0.67	1.00	0.18	0.50	0.36	0.42	0.75
TER	0.37	0.73	1.00	0.15	0.46	0.31	0.29	0.73
<b>UCI Dataset</b>								
Degree	0.48	0.29	0.98	0.44	0.28	0.26	0.32	0.44
PageRank	0.51	0.33	0.98	0.45	0.33	0.26	0.34	0.48
Cosine	0.51	0.34	0.33	0.67	0.46	0.51	0.59	0.82
Jaccard	0.57	0.36	0.28	0.84	0.49	0.45	0.65	0.84
TER	0.63	0.62	0.26	1.00	0.52	0.37	0.44	0.67

Table 2: Anomaly-class precision (P) and recall (R) of anomaly detection methods after perturbed using LORETTA

When adversaries employ data poisoning techniques to degrade model performance, anomaly detection systems can be deployed to identify adversarial edges and effectively clean the dataset before training. Consequently, it is critical for poisoning attacks to remain undetectable to such systems. To demonstrate LORETTA’s stealth capabilities, we evaluate its robustness against 4 SotA edge-stream anomaly detection methods: MIDAS (Bhatia et al. 2020), F-FADE (Chang et al. 2021), AnoEdge-L, and AnoEdge-G (Bhatia et al. 2023). We focus on unsupervised approaches due to their practical relevance, as supervised methods require labeled anomaly datasets that are typically unavailable in real-world deployment scenarios.

Table 2 presents the precision and recall scores of anomaly detection methods on the adversarial edges class when LORETTA perturbs 30% of the training data for the TGN

Method	Wikipedia				UCI			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
Cosine	57.16 ± 0.77	58.90 ± 0.85	58.00 ± 0.91	58.25 ± 1.25	33.58 ± 1.58	35.84 ± 1.68	33.42 ± 1.78	34.27 ± 2.40
Random	62.61 ± 0.81	63.17 ± 1.26	62.61 ± 1.13	62.61 ± 0.47	35.06 ± 1.45	34.83 ± 2.00	34.60 ± 2.52	35.32 ± 1.34
Jaccard	58.56 ± 0.82	58.45 ± 0.79	58.60 ± 0.51	59.97 ± 1.47	35.54 ± 1.66	36.39 ± 1.68	32.69 ± 1.53	33.95 ± 1.76
Pagerank	61.18 ± 0.69	64.43 ± 0.42	64.91 ± 1.33	67.40 ± 0.89	35.01 ± 2.32	35.83 ± 1.87	32.33 ± 2.62	36.10 ± 0.65
Degree	60.78 ± 0.70	64.40 ± 0.53	65.11 ± 0.31	67.06 ± 1.01	34.64 ± 2.67	35.11 ± 1.85	32.21 ± 1.86	35.38 ± 1.45
TER	60.78 ± 0.63	63.04 ± 1.04	65.93 ± 0.51	66.78 ± 1.48	34.74 ± 2.54	33.57 ± 2.47	32.53 ± 1.83	36.36 ± 0.86

Table 3: Performance (%) of different removal strategies under varying levels of adversary’s knowledge on the Wikipedia and UCI datasets. Each value is reported as mean ± standard deviation over 5 runs.

model across Wikipedia and UCI datasets. Each anomaly detection system assigns a score to every edge, where higher scores indicate greater likelihood of being anomalous. We employ Youden’s thresholding method (Youden 1950) to determine the optimal threshold for anomaly classification, as it minimizes both false positives and false negatives.

The results reveal a fundamental trade-off inherent in anomaly detection systems. High precision with low recall indicates accurate identification of true anomalies but at the cost of allowing many perturbed edges to remain undetected. Conversely, high recall with low precision typically results in numerous false positives that erroneously classify legitimate edges as anomalies, potentially degrading model performance when removed, as observed in §7.2. Critically, across all dataset-attack-detector combinations, no method achieves both precision and recall scores exceeding 0.7 simultaneously. This demonstrates that LORETTA successfully evades SotA anomaly detection systems and remains practically undetectable in realistic deployment scenarios, establishing its effectiveness as a stealthy adversarial attack. We also report in Appendix A.18 the AUPRC score (area under the precision–recall curve) for each anomaly detector, i.e., a threshold-free summary that integrates performance across all decision thresholds.

#### 7.4 Ablation Studies

In this section, we analyze the sensitivity of LORETTA to  $p$  and the knowledge of the training dataset.

**Impact of the Knowledge of Adversary:** Table 3 demonstrates the effect of increasing knowledge and access to the training dataset to the adversary on TGN performance degradation using LORETTA on Wikipedia. We discover that increasing adversarial knowledge does not always lead to more effective attacks. Heuristic strategies like pagerank and degree underperform—even degrade—with more knowledge. These methods remove high-centrality nodes assuming they’re critical, but often eliminate redundant or noisy nodes, unintentionally improving performance. This suggests that vulnerabilities are concentrated in a small set of influential nodes; once these are removed, additional knowledge yields diminishing or negative returns.

Interestingly, random removal shows stable performance across knowledge levels, suggesting non-monotonicity in attack impact. The system’s response is highly non-linear; only specific perturbations degrade performance. These findings indicate that attack effectiveness depends on targeting specific vulnerable components rather than simply having more

information about the system. For the adversary knowledge ablation, we perturb up to 30% of the edges to study the effect of strong adversarial influence across strategies. We experimentally validate our hypothesis in Appendix A.16.

**Impact of Perturbation Rate:** Figure 3 demonstrates the effect of increasing  $p$  on TGN performance degradation using LORETTA on Wikipedia, compared against T-spear. Performance degrades steadily with increasing  $p$  before plateauing, indicating that LORETTA’s deterministic sparsification algorithm targets influential edges first. Beyond a certain threshold, removing additional edges yields diminishing returns as the algorithm begins targeting less critical connections that contribute minimally to model learning. Notably, LORETTA consistently outperforms the SotA baseline across all  $p$ , demonstrating superior effectiveness despite operating as a surrogate-free attack.

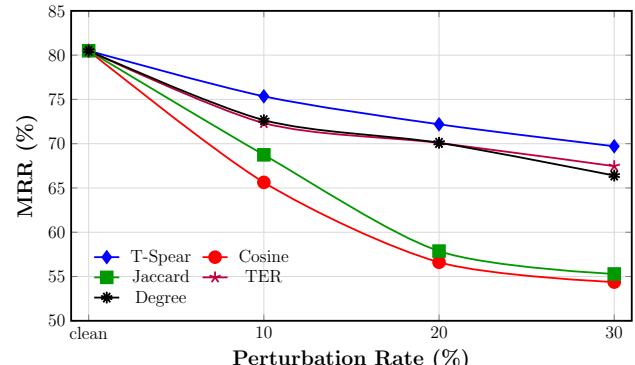


Figure 3: Effect of perturbation rate  $p$  on the performance of LORETTA using different sparsification strategies.

## 8 Conclusion

In this paper, we propose LORETTA, a novel attack strategy for TGNNs that exploits temporally significant edges to achieve effective perturbations. Our method consistently outperforms baselines, including T-SPEAR, across various datasets and models, highlighting the critical role of temporal dynamics in graph vulnerabilities. LORETTA is also robust to SotA adversarial defenses and anomaly detection systems.

## References

- Bhatia, S.; Hooi, B.; Yoon, M.; Shin, K.; and Faloutsos, C. 2020. Midas: Microcluster-based Detector of Anomalies in

- Edge Streams. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3242–3249.
- Bhatia, S.; Wadhwa, M.; Kawaguchi, K.; Shah, N.; Yu, P. S.; and Hooi, B. 2023. Sketch-based Anomaly Detection in Streaming Graphs. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 93–104.
- Bojchevski, A.; and Günnemann, S. 2019. Adversarial Attacks on Node Embeddings via Graph Poisoning. In *International Conference on Machine Learning*, 695–704.
- Cai, B.; Xiang, Y.; Gao, L.; Zhang, H.; Li, Y.; and Li, J. 2022. Temporal Knowledge Graph Completion: A Survey. *arXiv preprint:2201.08236*.
- Cencetti, G.; Santin, G.; Longa, A.; Pigani, E.; Barrat, A.; Cattuto, C.; Lehmann, S.; Salathe, M.; and Lepri, B. 2021. Digital Proximity Tracing on Empirical Contact Networks for Pandemic Control. *Nature communications*, 12(1): 1655.
- Chang, Y.-Y.; Li, P.; Sosic, R.; Afifi, M.; Schweighauser, M.; and Leskovec, J. 2021. F-fade: Frequency Factorization for Anomaly Detection in Edge Streams. In *Proceedings of the 14th ACM international conference on web search and data mining*, 589–597.
- Chen, J.; Zhang, J.; Chen, Z.; Du, M.; and Xuan, Q. 2021. Time-Aware Gradient Attack on Dynamic Network Link Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35: 2091–2102.
- Cong, W.; Zhang, S.; Kang, J.; Yuan, B.; Wu, H.; Zhou, X.; Tong, H.; and Mahdavi, M. 2023. Do We Really Need Complicated Model Architectures for Temporal Networks? *arXiv preprint:2302.11636*.
- Dai, H.; Li, H.; Tian, T.; Huang, X.; Wang, L.; Zhu, J.; and Song, L. 2018. Adversarial Attack on Graph Structured Data. In *International Conference on Machine Learning*, 1115–1124.
- Deng, S.; Rangwala, H.; and Ning, Y. 2019. Learning Dynamic Context Graphs for Predicting Social Events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1007–1016.
- Ennadir, S.; Gandler, G. Z.; Cornell, F.; Cao, L.; Smirnov, O.; Wang, T.; Zólyomi, L.; Brinne, B.; and Asadi, S. 2024. Expressivity of Representation Learning on Continuous-Time Dynamic Graphs: An Information-Flow Centric Review. *arXiv:2412.03783*.
- Entezari, N.; Al-Sayouri, S. A.; Darvishzadeh, A.; and Pa-palexakis, E. E. 2020. All you need is low (rank) defending against adversarial attacks on graphs. In *Proceedings of the 13th international conference on web search and data mining*, 169–177.
- Fan, W.; Ma, Y.; Li, Q.; He, Y.; Zhao, E.; Tang, J.; and Yin, D. 2019. Graph Neural Networks for Social Recommendation. In *The World Wide Web Conference*, 417–426.
- Feng, W.; Tang, J.; and Liu, T. X. 2019. Understanding Dropouts in Moocs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 517–524.
- Gao, C.; Wang, X.; He, X.; and Li, Y. 2022. Graph Neural Networks for Recommender System. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, 1623–1625.
- Hussain, H.; Duricic, T.; Lex, E.; Helic, D.; Strohmaier, M.; and Kern, R. 2021. Structack: Structure-Based Adversarial Attacks on Graph Neural Networks. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 111–120.
- Jiang, W.; and Luo, J. 2022. Graph Neural Network for Traffic Forecasting: A Survey. *Expert Systems with Applications*, 207: 117921.
- Kingma, D. P. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint:1412.6980*.
- Kleitman, D.; and Wang, D. 1973. Algorithms for constructing graphs and digraphs with given valences and factors. *Discrete Mathematics*, 6(1): 79–88.
- Kucharczuk, N.; Wąs, T.; and Skibski, O. 2022. Pagerank for Edges: Axiomatic Characterization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5108–5115.
- Kumar, S.; Zhang, X.; and Leskovec, J. 2019. Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1269–1278.
- Lee, D.; Lee, J.; and Shin, K. 2024. Spear and Shield: Adversarial Attacks and Defense Methods for Model-Based Link Prediction on Continuous-Time Dynamic Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13374–13382.
- Li, J.; Zhang, H.; Han, Z.; Rong, Y.; Cheng, H.; and Huang, J. 2020. Adversarial Attack on Community Detection by Hiding Individuals. In *Proceedings of the web conference 2020*, 917–927.
- Liben-Nowell, D.; and Kleinberg, J. 2003. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, 556–559.
- Ma, J.; Ding, S.; and Mei, Q. 2020a. Towards More Practical Adversarial Attacks on Graph Neural Networks. *Advances in Neural Information Processing Systems*, 33: 4756–4766.
- Ma, J.; Ding, S.; and Mei, Q. 2020b. Towards More Practical Adversarial Attacks on Graph Neural Networks. *arXiv:2006.05057*.
- Ma, Y.; Guo, Z.; Ren, Z.; Tang, J.; and Yin, D. 2020. Streaming Graph Neural Networks. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, 719–728.
- Panzarasa, P.; Opsahl, T.; and Carley, K. M. 2009. Patterns and Dynamics of Users’ Behavior and Interaction: Network Analysis of an Online Community. *Journal of the American Society for Information Science and Technology*, 60: 911–932.
- Pareja, A.; Domeniconi, G.; Chen, J.; Ma, T.; Suzumura, T.; Kanezashi, H.; Kaler, T.; Schardl, T.; and Leiserson, C. 2020. Evolvegcn: Evolving Graph Convolutional Networks for Dynamic Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5363–5370.

- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Rossi, E.; Chamberlain, B.; Frasca, F.; Eynard, D.; Monti, F.; and Bronstein, M. 2020. Temporal Graph Networks for Deep Learning on Dynamic Graphs. *arXiv preprint:2006.10637*.
- Rozenshtein, P.; and Gionis, A. 2016. Temporal PageRank. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 674–689.
- Sankar, A.; Wu, Y.; Gou, L.; Zhang, W.; and Yang, H. 2018. Dynamic Graph Representation Learning via Self-Attention Networks. *arXiv preprint:1812.09430*.
- Sankar, A.; Wu, Y.; Gou, L.; Zhang, W.; and Yang, H. 2020. Dysat: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 519–527.
- Sharma, K.; Trivedi, R.; Sridhar, R.; and Kumar, S. 2023. Temporal Dynamics-Aware Adversarial Attacks on Discrete-Time Dynamic Graph Models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2023–2035.
- Shetty, J.; and Adibi, J. 2004. The Enron Email Dataset Database Schema and Brief Statistical Report. *Information Sciences Institute Technical Report, University of Southern California*, 4: 120–128.
- So, M. K.; Tiwari, A.; Chu, A. M.; Tsang, J. T.; and Chan, J. N. 2020. Visualizing Covid-19 Pandemic Risk Through Network Connectedness. *International Journal of Infectious Diseases*, 96: 558–561.
- Trivedi, R.; Farajtabar, M.; Biswal, P.; and Zha, H. 2019. Dyrep: Learning Representations over Dynamic Graphs. In *International Conference on Learning Representations*.
- Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and Contributors, S. . 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. volume 17, 261–272.
- Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; et al. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.
- Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2022. Graph Neural Networks in Recommender Systems: A Survey. *ACM Computing Surveys*, 55(5): 1–37.
- Xu, D.; Ruan, C.; Korpeoglu, E.; Kumar, S.; and Acham, K. 2020. Inductive Representation Learning on Temporal Graphs. *arXiv preprint:2002.07962*.
- Youden, W. J. 1950. Index for Rating Diagnostic Tests. *Cancer*, 3(1): 32–35.
- Yu, B.; Yin, H.; and Zhu, Z. 2017. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. *arXiv preprint:1709.04875*.
- Zhang, X.; and Zitnik, M. 2020. Gnnguard: Defending graph neural networks against adversarial attacks. *Advances in neural information processing systems*, 33: 9263–9275.
- Zheng, Y.; Wei, Z.; and Liu, J. 2023. Decoupled graph neural networks for large dynamic graphs. *arXiv preprint arXiv:2305.08273*.
- Zhou, H.; Zheng, D.; Nisa, I.; Ioannidis, V.; Song, X.; and Karypis, G. 2022. Tgl: A General Framework for Temporal Gnn Training on Billion-Scale Graphs. *arXiv preprint:2203.14883*.
- Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial Attacks on Neural Networks for Graph Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2847–2856.
- Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial Attacks on Neural Networks for Graph Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’18. ACM.
- Zügner, D.; and Günnemann, S. 2024. Adversarial Attacks on Graph Neural Networks via Meta Learning. *arXiv:1902.08412*.

## A Appendix

### A.1 Detailed Explanation of Heuristics Used for Sparsification

We use deterministic sparsification strategies to remove a fixed fraction of edges from the temporal interaction graph. Each method scores edges or timestamps using a heuristic, then removes the top-ranked ones. The goal is to subtly weaken the model by pruning structurally or semantically important regions, all while ensuring reproducibility through sorted and tie-broken decisions.

Heuristics like Degree, Jaccard, PageRank, Preference, and Random (cf. §A.11) follow standard static scoring schemes. Here, we focus on sparsification methods built on two temporal graph signals: Temporal PageRank (TPR) and Temporal EdgeRank (TER).

**Temporal PageRank Drift Heuristics.** We track how node importance evolves over time by computing TPR scores at each timestamp. These scores are stored as a time-series vector for each node, forming a temporal embedding. To identify unstable or high-impact regions, we measure how much these vectors shift between timestamps using distance or divergence metrics.

- **Mean Shift in TPR (MSS & MSS<sup>2</sup>).** Both heuristics compute shifts in TPR scores across consecutive timestamps, but differ in the order of operations. In MSS, we first compute absolute per-node shifts and then take the mean. In MSS<sup>2</sup>, we compute the mean TPR vector across nodes at each timestamp, then take the shift between these mean vectors. The former emphasizes global volatility across nodes, while the latter highlights shifts in average importance.
- **Cosine Distance.** For each pair of consecutive timestamps, we compute the cosine distance between global TPR vectors:

$$1 - \frac{\vec{P}^{(t)} \cdot \vec{P}^{(t+1)}}{\|\vec{P}^{(t)}\|_2 \cdot \|\vec{P}^{(t+1)}\|_2}$$

This captures how the "direction" of node importance has changed, without being affected by absolute scale.

- **Jaccard Distance.** We binarize each TPR vector to retain the top- $k$  influential nodes at each timestamp, then compute:

$$1 - \frac{|B^{(t)} \cap B^{(t+1)}|}{|B^{(t)} \cup B^{(t+1)}|}$$

This captures how much the set of most important nodes changes over time.

- **Euclidean Distance.** This computes:

$$\|\vec{P}^{(t)} - \vec{P}^{(t+1)}\|_2$$

It measures the absolute magnitude of drift in node importance across timestamps.

- **Jensen–Shannon Divergence.** We normalize each TPR vector into a probability distribution and compute the symmetric divergence:

$$\text{JSD}(Q^{(t)} \| Q^{(t+1)}) = \frac{1}{2} \text{KL}(Q^{(t)} \| M) + \frac{1}{2} \text{KL}(Q^{(t+1)} \| M)$$

where  $M$  is the midpoint distribution. This gives a smooth, bounded signal for representational drift.

- **Kullback–Leibler Divergence.** Also operating on normalized TPR vectors, this asymmetric metric measures:

$$\text{KL}(Q^{(t)} \| Q^{(t+1)}) = \sum_i Q_i^{(t)} \log \frac{Q_i^{(t)}}{Q_i^{(t+1)} + \epsilon}$$

It emphasizes where a node's importance suddenly changes in one direction.

- **Chebyshev Distance.** This focuses on the node with the most drastic change in importance:

$$\max_i |P_i^{(t)} - P_i^{(t+1)}|$$

It surfaces sharp outliers in the node ranking space.

- **Wasserstein Distance.** Also known as Earth Mover's Distance, this computes how much "effort" it takes to transform one TPR distribution into the next. It's ideal for measuring both magnitude and position shifts in importance mass.

**Temporal EdgeRank (TER) Based Heuristics.** Unlike node-based TPR drift, these methods score and remove edges directly based on how influential they are in propagating information over time.

- **TER.** For each timestamp, we compute an importance score for every edge based on its temporal propagation potential. Edges with the highest TER are removed. This method directly targets structural bottlenecks in the flow of information.
- **Combined TER.** This heuristic blends two views of edge influence:
  - *Local*: Edge importance computed only up to timestamp  $t$
  - *Global*: Cumulative influence up to the final time  $T$

These are combined into a final score to capture both short-term volatility and long-term impact. The top-scoring edges across this combined metric are removed.

All pairwise distance computations (e.g., cosine, Euclidean, JSD, Wasserstein) over Temporal PageRank vectors were performed using the `scipy.spatial.distance` and `scipy.stats` modules from the SciPy library (Virtanen et al. 2020).

All the above methods aim to surface graph regions where representational dynamics or structural control are most active. The TPR-based methods take a node-centric view and quantify shifts in centrality embeddings. TER-based heuristics operate at edge-level, identifying links that most affect temporal signal flow. Together, they provide diverse ways of probing graph vulnerability.

### A.2 Attack Performance in Tabular Form

We show the Performance of 11 baselines and LORETTA on 4 TGNN victim models on 4 datasets in Table 4. We show the average of 5 runs.

Group	Attack	Model	Wikipedia	UCI	MOOC	Enron
Clean	No Attack	TGN	80.5% $\pm$ 0.5%	44.2% $\pm$ 0.4%	61.8% $\pm$ 2.1%	27.9% $\pm$ 0.4%
		JODIE	63.2% $\pm$ 1.2%	24.9% $\pm$ 0.4%	37.4% $\pm$ 2.0%	22.0% $\pm$ 0.4%
		DySAT	66.6% $\pm$ 0.4%	71.3% $\pm$ 0.9%	OOM	24.3% $\pm$ 0.7%
		TGAT	57.0% $\pm$ 0.6%	16.3% $\pm$ 0.3%	OOM	25.8% $\pm$ 0.6%
ADD Baselines	Degree	TGN	70.2% $\pm$ 0.7%	30.5% $\pm$ 1.5%	31.7% $\pm$ 0.5%	25.9% $\pm$ 0.7%
		JODIE	48.2% $\pm$ 1.8%	9.4% $\pm$ 0.3%	13.4% $\pm$ 0.6%	20.5% $\pm$ 1.5%
		DySAT	60.9% $\pm$ 0.5%	29.0% $\pm$ 1.4%	OOM	24.1% $\pm$ 0.2%
		TGAT	45.1% $\pm$ 0.6%	14.4% $\pm$ 0.5%	OOM	23.1% $\pm$ 0.2%
	Jaccard	TGN	—	30.3% $\pm$ 1.6%	—	25.4% $\pm$ 1.0%
		JODIE	—	11.1% $\pm$ 1.8%	—	18.6% $\pm$ 0.8%
		DySAT	—	29.7% $\pm$ 0.8%	—	23.6% $\pm$ 1.5%
		TGAT	—	11.9% $\pm$ 0.6%	—	22.8% $\pm$ 0.8%
REM Baselines	PageRank	TGN	69.9% $\pm$ 0.2%	28.9% $\pm$ 0.8%	31.1% $\pm$ 0.7%	26.6% $\pm$ 0.7%
		JODIE	48.0% $\pm$ 2.4%	10.1% $\pm$ 1.3%	13.4% $\pm$ 0.6%	26.7% $\pm$ 1.0%
		DySAT	60.4% $\pm$ 1.1%	29.5% $\pm$ 0.9%	OOM	19.6% $\pm$ 0.9%
		TGAT	44.6% $\pm$ 0.4%	15.7% $\pm$ 0.5%	OOM	23.0% $\pm$ 0.8%
	Preference	TGN	70.3% $\pm$ 1.1%	28.9% $\pm$ 1.2%	31.7% $\pm$ 1.2%	26.0% $\pm$ 0.5%
		JODIE	48.9% $\pm$ 2.0%	9.4% $\pm$ 0.3%	15.4% $\pm$ 0.6%	19.4% $\pm$ 0.4%
		DySAT	60.8% $\pm$ 0.8%	29.2% $\pm$ 1.2%	OOM	23.8% $\pm$ 0.8%
		TGAT	44.1% $\pm$ 0.5%	14.5% $\pm$ 0.3%	OOM	22.9% $\pm$ 0.5%
	Random	TGN	69.8% $\pm$ 0.4%	30.9% $\pm$ 1.2%	31.9% $\pm$ 0.4%	27.6% $\pm$ 0.7%
		JODIE	50.3% $\pm$ 2.0%	10.5% $\pm$ 1.3%	14.4% $\pm$ 0.5%	20.8% $\pm$ 0.3%
		DySAT	60.2% $\pm$ 1.1%	28.1% $\pm$ 1.9%	OOM	24.8% $\pm$ 1.1%
		TGAT	44.9% $\pm$ 0.8%	14.3% $\pm$ 0.4%	OOM	23.6% $\pm$ 0.2%
CTDG Poisoning Baseline	Degree	TGN	50.49% $\pm$ 1.35%	29.74% $\pm$ 0.56%	32.08% $\pm$ 0.21%	47.60% $\pm$ 0.62%
		JODIE	27.45% $\pm$ 0.96%	15.43% $\pm$ 0.49%	14.27% $\pm$ 0.44%	45.79% $\pm$ 0.60%
		DySAT	42.81% $\pm$ 1.00%	31.80% $\pm$ 0.51%	OOM	44.02% $\pm$ 0.14%
		TGAT	55.54% $\pm$ 0.87%	24.59% $\pm$ 1.20%	OOM	45.95% $\pm$ 0.41%
	Jaccard	TGN	—	24.20% $\pm$ 1.04%	—	53.83% $\pm$ 0.82%
		JODIE	—	11.05% $\pm$ 0.72%	—	52.74% $\pm$ 0.86%
		DySAT	—	25.01% $\pm$ 0.81%	—	49.97% $\pm$ 0.47%
		TGAT	—	18.36% $\pm$ 0.89%	—	51.68% $\pm$ 0.20%
	Pagerank	TGN	53.02% $\pm$ 1.55%	27.91% $\pm$ 1.19%	31.80% $\pm$ 0.54%	47.64% $\pm$ 0.33%
		JODIE	29.93% $\pm$ 2.96%	14.42% $\pm$ 0.70%	14.73% $\pm$ 0.32%	45.94% $\pm$ 0.61%
		DySAT	46.13% $\pm$ 0.64%	29.62% $\pm$ 1.11%	OOM	44.90% $\pm$ 0.65%
		TGAT	60.72% $\pm$ 0.58%	24.01% $\pm$ 1.26%	OOM	46.43% $\pm$ 0.41%
	Preference	TGN	48.97% $\pm$ 1.12%	24.03% $\pm$ 1.71%	29.07% $\pm$ 0.58%	44.87% $\pm$ 0.47%
		JODIE	28.02% $\pm$ 0.97%	11.52% $\pm$ 0.77%	14.23% $\pm$ 0.36%	43.50% $\pm$ 0.66%
		DySAT	42.38% $\pm$ 1.27%	26.31% $\pm$ 1.24%	OOM	41.71% $\pm$ 0.38%
		TGAT	55.34% $\pm$ 1.12%	19.33% $\pm$ 0.76%	OOM	42.47% $\pm$ 0.45%
	Random	TGN	57.35% $\pm$ 1.68%	21.79% $\pm$ 1.24%	34.40% $\pm$ 0.17%	40.99% $\pm$ 0.26%
		JODIE	40.05% $\pm$ 1.59%	11.74% $\pm$ 0.67%	17.96% $\pm$ 0.34%	39.19% $\pm$ 0.80%
		DySAT	41.55% $\pm$ 1.03%	23.75% $\pm$ 0.67%	OOM	39.07% $\pm$ 0.21%
		TGAT	63.90% $\pm$ 0.89%	17.27% $\pm$ 0.96%	OOM	41.99% $\pm$ 0.35%
LoReTTA	T-spear	TGN	69.8% $\pm$ 1.1%	30.9% $\pm$ 1.0%	32.1% $\pm$ 0.6%	27.2% $\pm$ 0.4%
		JODIE	51.2% $\pm$ 1.9%	19.8% $\pm$ 0.8%	14.4% $\pm$ 0.3%	20.4% $\pm$ 0.7%
		DySAT	59.5% $\pm$ 0.7%	31.3% $\pm$ 1.3%	OOM	25.2% $\pm$ 1.7%
		TGAT	43.8% $\pm$ 0.9%	10.8% $\pm$ 0.7%	OOM	22.9% $\pm$ 0.5%
	Degree	TGN	66.15% $\pm$ 0.96%	17.51% $\pm$ 1.39%	20.11% $\pm$ 0.40%	12.31% $\pm$ 0.68%
		JODIE	31.35% $\pm$ 1.22%	9.97% $\pm$ 1.23%	11.46% $\pm$ 0.35%	9.55% $\pm$ 0.40%
		DySAT	51.78% $\pm$ 0.90%	19.65% $\pm$ 1.33%	OOM	10.40% $\pm$ 0.41%
		TGAT	54.48% $\pm$ 0.65%	10.60% $\pm$ 0.27%	OOM	10.78% $\pm$ 0.24%
	Pagerank	TGN	66.51% $\pm$ 1.22%	17.18% $\pm$ 0.96%	19.75% $\pm$ 0.77%	13.40% $\pm$ 0.47%
		JODIE	30.65% $\pm$ 2.21%	9.00% $\pm$ 0.59%	11.99% $\pm$ 0.34%	10.82% $\pm$ 0.23%
		DySAT	53.71% $\pm$ 1.35%	19.60% $\pm$ 1.14%	OOM	11.16% $\pm$ 0.34%
		TGAT	56.10% $\pm$ 0.80%	11.05% $\pm$ 0.59%	OOM	10.85% $\pm$ 0.50%
	LoReTTA – TER	TGN	67.47% $\pm$ 0.96%	15.46% $\pm$ 1.06%	30.70% $\pm$ 0.50%	19.00% $\pm$ 0.53%
		JODIE	33.86% $\pm$ 0.56%	9.21% $\pm$ 1.31%	14.43% $\pm$ 0.57%	16.97% $\pm$ 0.46%
		DySAT	52.88% $\pm$ 1.84%	18.98% $\pm$ 0.71%	OOM	18.44% $\pm$ 0.93%
		TGAT	61.46% $\pm$ 0.48%	10.76% $\pm$ 0.61%	OOM	20.21 $\pm$ 0.67%
	LoReTTA – Cosine	TGN	54.05% $\pm$ 0.81%	15.42% $\pm$ 1.24%	23.25% $\pm$ 0.61%	20.30% $\pm$ 0.42%
		JODIE	28.15% $\pm$ 0.54%	9.55% $\pm$ 1.09%	11.78% $\pm$ 0.62%	17.03% $\pm$ 0.41%
		DySAT	48.39% $\pm$ 1.66%	23.61% $\pm$ 0.89%	OOM	17.00% $\pm$ 0.11%
		TGAT	47.53% $\pm$ 0.86%	11.63% $\pm$ 0.47%	OOM	13.37% $\pm$ 0.77%
	LoReTTA – Jaccard	TGN	55.18% $\pm$ 0.42%	15.65% $\pm$ 0.48%	22.91% $\pm$ 0.70%	18.28% $\pm$ 0.32%
		JODIE	28.71% $\pm$ 1.72%	9.16% $\pm$ 0.37%	10.92% $\pm$ 0.65%	15.64% $\pm$ 0.72%
		DySAT	48.21% $\pm$ 1.26%	22.42% $\pm$ 1.01%	OOM	15.72% $\pm$ 0.41%
		TGAT	49.25% $\pm$ 0.69%	11.23% $\pm$ 0.31%	OOM	13.33% $\pm$ 0.50%

Table 4: Attack performance of 4 TGNN models on 4 datasets against all attack methods.

### A.3 Attack performance of all tested Sparsification heuristics of LORETTA

We show the performance of LORETTA with 16 different sparsification strategies on 4 TGNN models and 4 datasets in Table 5. In our sparsification framework, we compute the Temporal PageRank (TPR) vector  $r^{(t)} \in \mathbb{R}^{|V|}$  at each timestamp  $t$  to quantify evolving node influence. The temporal drift between two successive timestamps is measured by comparing  $r^{(t)}$  and  $r^{(t-1)}$  using a chosen similarity or divergence metric. Our empirical evaluation reveals that similarity-based metrics—particularly cosine similarity and Jaccard index—consistently outperform conventional distance-based metrics such as KL divergence, Jensen–Shannon (JS) divergence, and  $\ell_1$  or  $\ell_2$  norms. We now analyze the underlying reasons for this performance gap.

**Sparsity and high dimensionality in TPR vectors.** TPR vectors are typically high-dimensional and sparse since only a fraction of nodes exhibit temporal influence at each timestamp. In such regimes, distance-based metrics can become unstable or uninformative. For instance, KL and JS divergence assume overlapping supports and suffer when distributions contain zeros. In contrast, cosine and Jaccard similarity are tailored for sparse domains: cosine focuses on angular displacement (not magnitude), while Jaccard tracks support set overlap.

**Directional sensitivity of cosine similarity.** Cosine similarity captures shifts in influence direction across time:

$$\cos \theta = \frac{\langle r^{(t)}, r^{(t-1)} \rangle}{\|r^{(t)}\| \cdot \|r^{(t-1)}\|}.$$

This normalized, scale-invariant signal aligns with the model’s sensitivity to which nodes are gaining or losing importance, rather than their absolute rank. TGNNs are particularly vulnerable to such directional changes in influence dynamics, making cosine similarity an effective proxy for semantic drift.

**Support tracking via Jaccard index.** Jaccard similarity computes the ratio of shared active nodes between consecutive TPR vectors:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

$$A = \text{support}(r^{(t)}),$$

$$B = \text{support}(r^{(t-1)})$$

This metric directly reflects reconfiguration in the set of influential nodes, which is highly correlated with performance drops in TGNNs trained on evolving interaction patterns. As sparsification aims to target such volatile moments, Jaccard serves as a reliable indicator of high-impact intervals.

**Latent semantic alignment and adversarial leverage.** Beyond their structural robustness, similarity metrics also operate in a space more aligned with the model’s predictive behavior. Cosine similarity, in particular, indirectly reflects alignment in the TGNN’s latent embedding space—capturing semantic closeness between nodes. Removal of edges between

such semantically aligned nodes degrades the coherence of learned prototypes, leading to impaired message passing and degraded representation learning.

**Representation space is more fragile than observable structure.** Our findings suggest that TGNNs are more sensitive to latent-space perturbations than to explicit topological alterations. While structural heuristics such as node degree or temporal frequency affect global statistics, similarity-based metrics destabilize the representation space in ways that are more difficult to recover from—especially under limited-knowledge or surrogate-free attack settings.

In sum, similarity-based metrics by virtue of their alignment with both sparsity patterns and semantic drift serve as more effective tools for perturbation planning in temporal graphs. Their ability to implicitly target the model’s latent structure enables stronger degradation with weaker assumptions, establishing them as superior alternatives to classical distance measures in the context of TGNN adversarial attacks.

### A.4 Compliance with C3 and C4 Constraints

We empirically validate that our adversarial edge generation procedure respects the **C3 (recent node activity)** and **C4 (node degree preservation)** constraints, both of which are critical for stealth in continuous-time poisoning attacks.

**C3: Temporal Activity Window.** Constraint C3 ensures that any adversarial edge connects nodes that were recently active prior to the assigned timestamp. To enforce this, our EdgeTimestamp Selector explicitly filters candidates based on temporal recency. Figure 4 shows the timestamp-node activity alignment on the Wikipedia dataset. All inserted edges lie within the expected window of recent activity, confirming our compliance with C3 and the temporal plausibility of generated interactions.

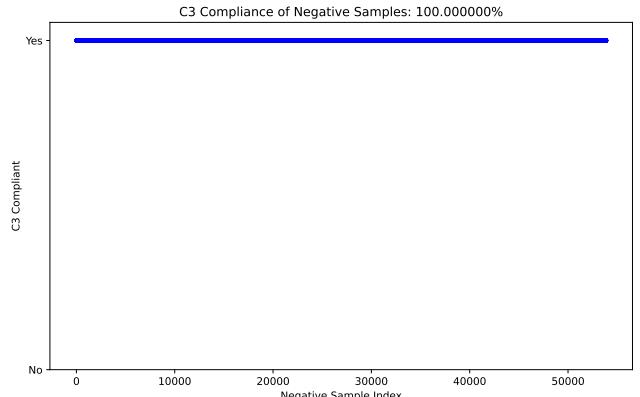
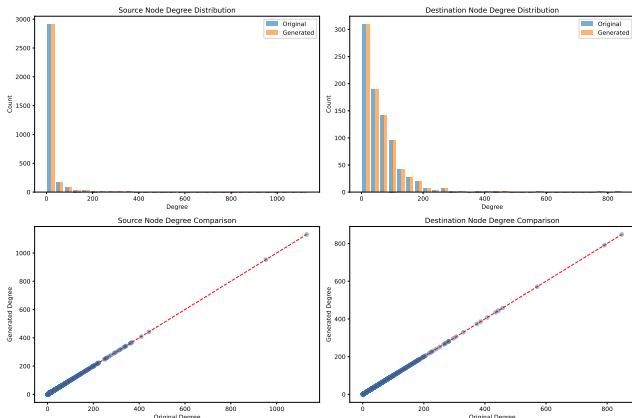


Figure 4: Compliance with the C3 constraint (Wikipedia dataset). Each adversarial edge connects nodes that were active in a fixed window prior to the sampled timestamp.

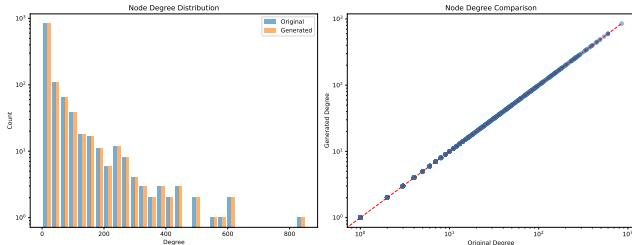
**C4: Degree Preservation.** Constraint C4 aims to preserve node degree distributions post-attack, thereby maintaining structural stealth. We track insertion and deletion counts

Attack	Model	Wikipedia	UCI	MOOC	Enron
Degree	TGN	66.15% $\pm$ 0.96%	17.51% $\pm$ 1.39%	20.11% $\pm$ 0.40%	12.31% $\pm$ 0.68%
	JODIE	31.35% $\pm$ 1.22%	9.97% $\pm$ 1.23%	11.46% $\pm$ 0.35%	9.55% $\pm$ 0.40%
	DySAT	51.78% $\pm$ 0.90%	19.65% $\pm$ 1.33%	–	10.40% $\pm$ 0.41%
	TGAT	54.48% $\pm$ 0.65%	10.60% $\pm$ 0.27%	–	–
Jaccard	TGN	–	18.63% $\pm$ 0.77%	–	19.28% $\pm$ 0.51%
	JODIE	–	10.27% $\pm$ 0.90%	–	16.43% $\pm$ 0.17%
	DySAT	–	22.14% $\pm$ 1.22%	–	17.46% $\pm$ 0.14%
	TGAT	–	11.15% $\pm$ 0.19%	–	–
Pagerank	TGN	66.51% $\pm$ 1.22%	17.18% $\pm$ 0.96%	19.75% $\pm$ 0.77%	13.40% $\pm$ 0.47%
	JODIE	30.65% $\pm$ 2.21%	9.00% $\pm$ 0.59%	11.99% $\pm$ 0.34%	10.82% $\pm$ 0.23%
	DySAT	53.71% $\pm$ 1.35%	19.60% $\pm$ 1.14%	–	11.16% $\pm$ 0.34%
	TGAT	56.10% $\pm$ 0.80%	11.05% $\pm$ 0.59%	–	–
Preference	TGN	65.25% $\pm$ 1.46%	17.42% $\pm$ 1.61%	25.70% $\pm$ 0.33%	14.25% $\pm$ 0.39%
	JODIE	27.85% $\pm$ 3.76%	9.93% $\pm$ 1.66%	12.28% $\pm$ 0.73%	10.11% $\pm$ 0.49%
	DySAT	52.75% $\pm$ 1.50%	22.02% $\pm$ 1.50%	–	11.75% $\pm$ 0.18%
	TGAT	55.10% $\pm$ 0.26%	10.59% $\pm$ 0.45%	–	–
Random	TGN	62.06% $\pm$ 1.35%	17.40% $\pm$ 1.49%	29.09% $\pm$ 0.64%	27.46% $\pm$ 0.69%
	JODIE	24.18% $\pm$ 0.72%	9.36% $\pm$ 1.18%	12.36% $\pm$ 0.38%	25.09% $\pm$ 1.29%
	DySAT	48.14% $\pm$ 1.68%	22.82% $\pm$ 1.55%	–	25.66% $\pm$ 0.66%
	TGAT	55.38% $\pm$ 1.12%	10.70% $\pm$ 0.34%	–	–
TER	TGN	67.47% $\pm$ 0.96%	15.46% $\pm$ 1.06%	30.70% $\pm$ 0.50%	19.00% $\pm$ 0.53%
	JODIE	33.86% $\pm$ 0.56%	9.21% $\pm$ 1.31%	14.43% $\pm$ 0.57%	16.97% $\pm$ 0.46%
	DySAT	52.88% $\pm$ 1.84%	18.98% $\pm$ 0.71%	–	18.44% $\pm$ 0.93%
	TGAT	61.46% $\pm$ 0.48%	10.76% $\pm$ 0.61%	–	–
TPR-Chebyshev	TGN	68.67% $\pm$ 0.56%	18.05% $\pm$ 0.83%	33.21% $\pm$ 0.95%	20.18% $\pm$ 1.08%
	JODIE	41.38% $\pm$ 1.69%	9.88% $\pm$ 0.64%	16.44% $\pm$ 0.62%	18.28% $\pm$ 0.70%
	DySAT	53.28% $\pm$ 1.47%	21.43% $\pm$ 0.94%	–	19.50% $\pm$ 1.25%
	TGAT	63.50% $\pm$ 0.72%	11.22% $\pm$ 0.33%	–	–
Combined-TER	TGN	68.42% $\pm$ 0.74%	17.40% $\pm$ 1.03%	33.39% $\pm$ 1.43%	21.36% $\pm$ 0.41%
	JODIE	40.71% $\pm$ 2.41%	10.19% $\pm$ 1.43%	16.58% $\pm$ 0.63%	19.23% $\pm$ 0.83%
	DySAT	53.82% $\pm$ 1.21%	19.81% $\pm$ 0.71%	–	19.26% $\pm$ 1.01%
	TGAT	63.70% $\pm$ 0.64%	10.97% $\pm$ 0.27%	–	–
TPR-Cosine	TGN	54.05% $\pm$ 0.81%	15.42% $\pm$ 1.24%	23.25% $\pm$ 0.61%	20.30% $\pm$ 0.42%
	JODIE	28.15% $\pm$ 0.54%	9.55% $\pm$ 1.09%	11.78% $\pm$ 0.62%	17.03% $\pm$ 0.41%
	DySAT	48.39% $\pm$ 1.66%	23.61% $\pm$ 0.89%	–	17.00% $\pm$ 0.11%
	TGAT	47.53% $\pm$ 0.86%	11.63% $\pm$ 0.47%	–	16.29% $\pm$ 0.55%
TPR-Euclidean	TGN	68.63% $\pm$ 1.08%	17.93% $\pm$ 0.84%	33.23% $\pm$ 0.80%	21.63% $\pm$ 0.63%
	JODIE	40.79% $\pm$ 1.98%	10.83% $\pm$ 1.43%	16.93% $\pm$ 0.73%	18.21% $\pm$ 1.08%
	DySAT	53.45% $\pm$ 1.10%	22.95% $\pm$ 1.19%	–	18.84% $\pm$ 0.79%
	TGAT	63.49% $\pm$ 0.81%	11.21% $\pm$ 0.44%	–	24.31% $\pm$ 0.50%
TPR-Jaccard	TGN	55.18% $\pm$ 0.42%	15.65% $\pm$ 0.48%	22.91% $\pm$ 0.70%	18.28% $\pm$ 0.32%
	JODIE	28.71% $\pm$ 1.72%	9.16% $\pm$ 0.37%	10.92% $\pm$ 0.65%	15.64% $\pm$ 0.72%
	DySAT	48.21% $\pm$ 1.26%	22.42% $\pm$ 1.01%	–	15.72% $\pm$ 0.41%
	TGAT	49.25% $\pm$ 0.69%	11.23% $\pm$ 0.31%	–	–
TPR-Jensen Shannon Divergence	TGN	66.23% $\pm$ 0.34%	19.06% $\pm$ 1.31%	32.90% $\pm$ 0.55%	16.61% $\pm$ 0.70%
	JODIE	34.19% $\pm$ 1.14%	12.12% $\pm$ 1.46%	15.89% $\pm$ 0.87%	15.07% $\pm$ 0.84%
	DySAT	50.65% $\pm$ 0.53%	22.27% $\pm$ 1.19%	–	16.44% $\pm$ 0.78%
	TGAT	61.40% $\pm$ 0.56%	11.39% $\pm$ 0.28%	–	–
TPR-KL Divergence	TGN	66.15% $\pm$ 0.83%	19.93% $\pm$ 1.71%	32.64% $\pm$ 0.66%	16.73% $\pm$ 0.73%
	JODIE	35.61% $\pm$ 1.36%	11.40% $\pm$ 1.21%	15.06% $\pm$ 1.16%	15.12% $\pm$ 0.51%
	DySAT	51.52% $\pm$ 1.30%	22.77% $\pm$ 0.87%	–	16.53% $\pm$ 0.72%
	TGAT	61.24% $\pm$ 0.87%	11.60% $\pm$ 0.29%	–	–
TPR-MSS	TGN	68.70% $\pm$ 1.00%	18.82% $\pm$ 1.89%	33.30% $\pm$ 0.78%	22.21% $\pm$ 0.69%
	JODIE	40.20% $\pm$ 1.10%	10.97% $\pm$ 1.46%	15.16% $\pm$ 1.67%	19.00% $\pm$ 0.85%
	DySAT	53.04% $\pm$ 1.53%	22.98% $\pm$ 1.04%	–	19.12% $\pm$ 1.67%
	TGAT	64.16% $\pm$ 0.35%	11.59% $\pm$ 0.25%	–	24.28% $\pm$ 0.66%
TPR-MSS <sup>2</sup>	TGN	68.45% $\pm$ 1.11%	18.64% $\pm$ 1.16%	33.43% $\pm$ 0.24%	21.69% $\pm$ 0.76%
	JODIE	40.27% $\pm$ 1.93%	10.19% $\pm$ 1.32%	15.88% $\pm$ 2.13%	18.80% $\pm$ 0.37%
	DySAT	52.78% $\pm$ 0.81%	21.41% $\pm$ 0.95%	–	19.14% $\pm$ 1.71%
	TGAT	63.50% $\pm$ 0.59%	11.30% $\pm$ 0.45%	–	–
TPR-Wasserstein	TGN	67.72% $\pm$ 0.67%	17.74% $\pm$ 0.70%	33.60% $\pm$ 0.94%	21.71% $\pm$ 0.58%
	JODIE	39.28% $\pm$ 1.66%	9.66% $\pm$ 1.16%	15.87% $\pm$ 2.28%	18.26% $\pm$ 0.97%
	DySAT	53.28% $\pm$ 1.32%	22.24% $\pm$ 1.31%	–	19.53% $\pm$ 1.17%
	TGAT	63.70% $\pm$ 0.56%	11.21% $\pm$ 0.37%	–	–

Table 5: Test MRR (%) of LoRETTA using various sparsification strategies across 4 datasets and 4 TGNs.



(a) Bipartite dataset. Red dashes show original degrees; blue dots denote regenerated degrees. High overlap confirms C4 compliance.



(b) Non-bipartite dataset (UCI). Degree preservation is evident through strong alignment of regenerated and original degrees.

Figure 5: C4 constraint validation across different graph types. Regenerated degrees preserve structural statistics, ensuring stealth.

per node, ensuring that adversarial edges compensate for removed ones in a degree-balanced fashion.

Figures 5a and 5b show degree distributions for bipartite and non-bipartite settings respectively. Both histograms and scatter plots confirm that regenerated node degrees closely mirror the original graph, validating the effectiveness of our degree-preserving strategy. In particular, the scatter plots reveal a near one-to-one match between original and modified degrees.

**Takeaway.** Together, these results demonstrate that our negative sampling algorithm is *constraint-aware by design*. It satisfies both temporal realism (C3) and structural stealth (C4), enabling a practical and undetectable attack strategy for CTDGs.

## A.5 Time and Space Complexity Analysis

We analyze both the theoretical and empirical complexity of LORETTA, and show that it is significantly more efficient than T-SPEAR (Lee, Lee, and Shin 2024), the current state-of-the-art.

### Theoretical Time Complexity

**Lemma 1** (Temporal PageRank Runtime). *Let  $G = (V, E)$  be a temporal graph, and  $\alpha, \beta \in [0, 1]$  be the jump and decay*

*parameters. Algorithm TPR runs in*

$$O(|E| + |V|) \text{ time.}$$

*Proof.* We maintain vectors  $r, s : V \rightarrow \mathbb{R}_{\geq 0}$  to track scores and intermediate contributions. Each edge  $(u, v, t)$  is processed using a constant-time update map  $F_{\alpha, \beta}$ , which requires  $O(1)$  operations. Thus, processing all edges takes  $O(|E|)$  time.

Afterward, a normalization pass over all vertices takes  $O(|V|)$  time. Total cost:  $O(|E| + |V|)$ .  $\square$

**Lemma 2** (Timestamp Selector Runtime). *Let  $d_{\max}$  be the maximum node degree and  $k$  be the average number of candidate timestamps per edge. The TIMESTAMP SELECTOR algorithm runs in*

$$O(|V|(\log |V| + d_{\max}k \log k) + |E| \log |E|).$$

*Proof.* (i) **Initialization:** Sorting vertices takes  $O(|V| \log |V|)$ ; computing edge budgets takes  $O(|E|)$ .

(ii) **Edge-timestamp selection:** For each node  $v \in V$ , we:

- Check capacity:  $O(1)$ ,
- Collect candidate edges:  $O(d_{\max})$ ,
- Select timestamps:  $O(k \log k)$ .

Total:  $O(|V|d_{\max}k \log k)$ .

(iii) **Recovery (if needed):** Sorting edges dominates with  $O(|E| \log |E|)$ .

Combining phases yields the final bound.  $\square$

**Empirical Time Complexity** We evaluate runtime against T-SPEAR across four benchmark datasets using TGN as the base TGNN and a fixed 30% perturbation rate.

Table 6: Wall-clock time (in seconds) for LORETTA and T-SPEAR. Lower is better.

Method	Wikipedia	Enron	UCI	MOOC
T-SPEAR	4299.02	2198.16	996.26	7320.78
LORETTA	370.18	620.30	434.36	1595.43

LORETTA achieves up to a  $10\times$  speedup over T-SPEAR, with an average of  $3.91\times$  lower runtime across all datasets. This includes:

- Surrogate model training (only for T-SPEAR),
- Edge addition (both),
- Edge removal (only for LORETTA).

In contrast to T-SPEAR, which relies on training a learned surrogate model and computing model gradients for edge ranking, LORETTA operates purely via heuristics over the interaction graph. This removes the need for backpropagation, reduces memory overhead, and enables black-box deployment. **As a result, LORETTA is significantly more resource-efficient while maintaining strong attack performance.**

## Space Complexity

**Lemma 3** (TPR and TER Memory Footprint). *Let  $|V|$  be the number of nodes and  $|T|$  the number of unique timestamps. The memory used by TPR and TER is*

$$O(|V||T|).$$

*Proof.* We store one PageRank vector  $r_t : V \rightarrow \mathbb{R}_{\geq 0}$  per timestamp  $t$ , totaling  $|V||T|$  scalars. An auxiliary degree-tracking array adds  $O(|V|)$  overhead.  $\square$

**Lemma 4** (Timestamp Selector Memory Footprint). *Let  $W$  be the time window for candidate selection. The TIMESTAMP SELECTOR uses*

$$O(|V| + |E|W)$$

space.

*Proof.* Base graph storage requires  $O(|V| + |E|)$ . Each edge may generate up to  $W$  candidates, resulting in  $O(|E|W)$  auxiliary storage. Additional counters per vertex require  $O(|V|)$ .  $\square$

## A.6 Implementation Details

For all experiments, we utilized a constant learning rate of 0.0001 across the training of victim models, surrogate models, attack methods, and defense models. The Adam optimizer (Kingma 2014) was employed with a batch size of 600, and all models were trained for a total of 100 epochs (50 epochs for DySAT). An early stopping criterion was applied after 50 epochs (25 epochs for DySAT), where training was halted if the validation Mean Reciprocal Rank (MRR) did not improve for 10 consecutive epochs.

## A.7 Hyperparameters for Models

The models, including TGN (Rossi et al. 2020), JODIE (Kumar, Zhang, and Leskovec 2019), TGAT (Xu et al. 2020), and DySAT (Sankar et al. 2020), were implemented using the Temporal Graph Learning<sup>1</sup> (TGL) framework (Zhou et al. 2022). This framework provides optimized implementations, achieving better overall scores compared to the original implementations of the respective models. Consequently, we adopted TGL’s default hyperparameter settings for all victim models across datasets. These settings include a learning rate of 0.0001, batch size of 600, and a hidden dimension of 100. A detailed summary of hyperparameters is provided in Table 7.

## A.8 Hardware Specifications and Environment

All experiments are conducted on one NVIDIA A100 GPU (40GB memory), 512GB of RAM, and AMD EPYC 7742 64-Core Processors. Our models are implemented using PyTorch 1.12.1 (Paszke et al. 2019) and Deep Graph Library (DGL) 0.9.1 (Wang et al. 2019).

<sup>1</sup>The TGL framework can be found at <https://github.com/amazonresearch/tgl>

Table 7: Hyperparameter Settings for Victim Models

Parameter	Value	Applicable Models
Learning Rate	0.0001	All
Batch Size	600	All
Hidden Dimension	100	All
Number of Epochs	100	All except DySAT
	50	DySAT
Early Stopping Patience	10 epochs	All

## A.9 Datasets

We use four datasets, which are publicly available <sup>2</sup>:

- **Wikipedia:** This is a bipartite interaction graph representing edits made on Wikipedia pages over a one-month period. Nodes correspond to users and pages, while links denote editing activities with associated timestamps. This dataset contains 157,474 attributed interactions recorded over one month, involving 8,227 users and 1,000 pages. Each interaction represents a user editing a page, with the editing texts converted into LIWC-feature vectors.
- **MOOC:** This dataset represents a bipartite interaction network of online educational resources, where nodes correspond to students and course content units (e.g., videos and problem sets). It comprises interactions from 7,047 users with 97 items, resulting in 411,749 attributed interactions recorded over approximately one month. These interactions represent the access behavior of students to online course units.
- **Enron:** This dataset captures email communications among employees of the ENRON energy corporation over a three-year period. Nodes represent employees, and links denote email exchanges with timestamps.
- **UCI:** This is an online communication network where nodes correspond to university students. Links represent messages exchanged between students, annotated with temporal information. It consists of 59,835 message interactions involving 1,899 unique users.

## A.10 Dataset Statistics

We conduct our experiments on four widely used temporal graph datasets spanning both bipartite and non-bipartite domains. Table 8 provides a summary.

Table 8: Summary of datasets used in experiments. Each dataset contains timestamped interactions and supports Unix-level granularity.

Dataset	Domain	Nodes	Edges	Bipartite	Duration	Granularity
Wikipedia	Social	9,227	157,474	Yes	1 Month	Unix Time
MOOC	Interaction	7,144	411,749	Yes	17 Months	Unix Time
UCI	Social	1,899	59,835	No	196 Days	Unix Time
Enron	Communication	184	125,235	No	3 Years	Unix Time

<sup>2</sup><https://zenodo.org/records/7213796#.Y1cO6y8r30o>

## A.11 Baselines

We benchmark LORETTA against a suite of heuristic baselines commonly used in link prediction and adversarial attack literature. These strategies generate perturbations either by adding new edges (used in prior works such as (Lee, Lee, and Shin 2024)) or by removing existing ones (used in our setting to highlight the effect of sparsification). Each method operates under the unnoticeability constraints outlined in §4, and metrics are computed using a plain graph constructed from edges available just prior to perturbation.

- **Random:** Adds or removes edges by uniformly sampling node pairs without replacement. Serves as a lower-bound baseline with no structural or temporal bias.
- **Preference:** Based on Preferential Attachment (PA) (Liben-Nowell and Kleinberg 2003), this method connects or deletes edges between nodes with the lowest PA score, where  $\text{PA}(u, v) = |N(u)| \cdot |N(v)|$ . Nodes with fewer connections are more likely to be perturbed, mimicking sparsity-aware behavior.
- **Jaccard:** Links or removes node pairs with the lowest Jaccard coefficient (Liben-Nowell and Kleinberg 2003), defined as

$$\text{Jaccard}(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}.$$

This baseline is applied only to unipartite graphs, as common neighbor sets are undefined in bipartite settings.

- **Degree:** Inspired by Structack (Hussain et al. 2021), this strategy targets node pairs with minimal degree sum, i.e.,  $|N(u)| + |N(v)|$ . It reflects a structural vulnerability perspective by focusing on low-connectivity regions.
- **PageRank:** Similar to Degree, this baseline perturbs edges between nodes with the lowest cumulative PageRank scores,  $PR(u) + PR(v)$ , targeting semantically weakly connected nodes.

For each baseline, we evaluate both edge addition (ADD) and edge removal (REM) variants: the former aligns with prior poisoning settings, while the latter is used in our context to demonstrate the effectiveness of edge sparsification. Together, these comparisons provide a comprehensive view of how simple graph heuristics fare against LORETTA’s temporal and constraint-aware attack strategy. **In the original paper (Lee, Lee, and Shin 2024), the entire dataset—including validation and test sets—is poisoned. In our setup, we poison only the training set to ensure a fair comparison with our method. This difference in setup explains the variation in reported numbers for the T-SPEAR attack between our paper and the original.**

## A.12 Defense Methods

These are the defense methods used.

1. **TGN-SVD:** Drawing inspiration from GCN-SVD (Entezari et al. 2020), this method constructs a plain graph using all edges from the training set and computes a low-rank approximation  $\hat{A}$  of the adjacency matrix. The entries of  $\hat{A}$  are used as weights in the loss function. For each

edge  $e = (u, v, t)$ , the weight of the edge’s loss in Eq. (1) is determined by  $\hat{a}_{uv}$ .

2. **TGN-COSINE:** Based on GNNGuard (Zhang and Zitnik 2020), this approach utilizes cosine similarity to filter edges. An edge  $e = (u, v, t)$  is removed if the cosine similarity between the node embeddings  $h_u(t)$  and  $h_v(t)$  falls below a predefined threshold  $\tau_{\text{cosine}}$ .
3. **TShield:** T-shield is proposed in Spear and Shield (Lee, Lee, and Shin 2024) and is one of the SotA defense methods for defense. It identifies and eliminates potential adversarial edges from the corrupted graphs, considering the evolution of dynamic graphs without prior knowledge of adversarial attacks.
4. **TShield-F:** T-Shield-F is a variant of the method proposed in Spear and Shield (Lee, Lee, and Shin 2024), retaining only the edge-filtering step while excluding the temporal smoothing step.

**Temporal Walk** In the context of temporal networks, a *temporal walk* (also referred to as a *time-respecting walk*) on a temporal graph  $G$  is defined as an ordered sequence of edges:

$$(u_1, u_2, t_1), (u_2, u_3, t_2), \dots, (u_j, u_{j+1}, t_j),$$

where each edge  $(u_i, u_{i+1}, t_i)$  connects nodes  $u_i$  and  $u_{i+1}$  at a timestamp  $t_i$ . A key property of a temporal walk is that the timestamps respect the order of the edges, i.e.,  $t_i \leq t_{i+1}$  for all  $1 \leq i \leq j - 1$ . This ensures that the walk adheres to the temporal constraints of the network.

## A.13 On the Stealthiness of LORETTA

Unlike node injection attacks that introduce new entities into the graph often detectable via anomaly detection or role analysis, LORETTA perturbs only interactions among existing, historically active nodes. This makes it significantly more stealthy. By adhering to unnoticeability constraints (C1–C4), the inserted edges mimic natural behavior in both temporal and structural ways. No new nodes are introduced, and edge insertions preserve degree and temporal-activity patterns, making them indistinguishable from genuine interactions (see A.4). This interaction-level perturbation enables practical and undetectable poisoning in real-world CTDGs, where node-level attacks would likely raise flags.

## A.14 Temporal PageRank and EdgeRank: Intuition, Formulation, and Computation

**Temporal PageRank** Temporal PageRank (TPR) extends classical PageRank to temporal graphs by replacing static walks with time-respecting walks (Rozenshtein and Giannis 2016). Let  $Z^T(v, u \mid t)$  denote the set of temporal walks (cf. §A.12) from node  $v$  to  $u$  that occur strictly before time  $t$ . The probability of a walk  $z$  is defined as:

$$\Pr[z \in Z^T(v, u \mid t)] = \frac{c(z \mid t)}{\sum_{z' \in Z^T(v, u \mid t), x \in V, |z'|=|z|} c(z' \mid t)},$$

where  $c(z | t)$  is the decay-weighted count of  $z$ :

$$c(z | t) = (1 - \beta) \prod_{\substack{((u_{i-1}, u_i, t_i), \\ (u_i, u_{i+1}, t_{i+1})) \in z}} \beta^{| \{(u_i, y, t') | t' \in [t_i, t_{i+1}], y \in V\} |}.$$

Transitions are penalized exponentially by the number of intervening interactions to model temporal decay. The final TPR score of node  $u$  at time  $t$  is:

$$r(u, t) = \sum_{v \in V} \sum_{k=0}^t (1 - \alpha) \alpha^k \sum_{\substack{z \in Z_T(v, u | t) \\ |z|=k}} \Pr[z | t],$$

where  $\alpha$  is the jump probability. This formulation biases the walk toward shorter, temporally coherent paths and naturally captures influence drift in evolving graphs.

**Temporal EdgeRank** To quantify the temporal importance of edges in the dynamic network, we introduce Temporal EdgeRank, which extends the concept of EdgeRank (Kucharczuk, Wąs, and Skibski 2022) to temporal graphs (cf. Algorithm 1). The algorithm first computes temporal PageRank scores for nodes using a damping factor  $\beta$  and restart probability  $\alpha$ , capturing how node importance evolves over time. For each timestamp  $t$ , the algorithm identifies the set of active nodes and calculates the EdgeRank score by normalizing each node's temporal PageRank value by its overall outgoing degree. This normalization ensures that the influence of highly connected nodes is appropriately scaled. Specifically, the Temporal EdgeRank score at time  $t$  is computed as the sum of the ratio between each active node's temporal PageRank score and its total outgoing degree. This approach effectively captures both the temporal dynamics of node importance and the structural properties of the network, providing a measure of edge significance that accounts for the temporal evolution of the graph structure. This approach integrates two complementary aspects of temporal graph analysis: (1) Temporal PageRank effectively captures the temporal dynamics of node importance by assigning higher weights to nodes participating in influential time-respecting walks, reflecting their centrality in evolving interactions. (2) Outgoing degree serves as a normalizing factor, quantifying the overall participation of a node across the temporal graph and preventing the measure from being biased solely by temporal activity bursts. Together, these components ensure that the significance of an edge reflects not just the structural properties of the graph at a single time point but also the temporal evolution of node roles and interactions over time.

Together, TPR and TER enable fine-grained reasoning about temporal dynamics in node and edge importance. They serve as the foundation for sparsification in our attack pipeline, helping identify structurally sensitive and temporally volatile regions of the graph.

**Negative Sampling Algorithm** The algorithm 2 employs a priority-based approach where nodes are first sorted based on their structural importance in the network. For each node in priority order, the algorithm selects valid edges (*GetValidEdges*) within a specified time window while maintaining

---

#### Algorithm 1: Temporal EdgeRank

---

**Input:** Edge set  $E$  as  $(u, v, t)$  tuples, damping  $\beta$ , restart prob.  $\alpha$   
**Output:** Temporal EdgeRank scores  
 $TPR \leftarrow \text{TemporalPageRank}(E, \beta, \alpha)$   
 $\text{out\_deg} \leftarrow \text{ComputeOutDegree}(E)$   
 $TER \leftarrow \{\}$  {Temporal EdgeRank}  
**for** all timestamp  $t$  in  $E$  **do**  
     $\text{node}_{st} \leftarrow \{u | (u, v, t) \in E\}$   
     $TER[t] \leftarrow \sum_{n \in \text{node}_{st}} \frac{TPR[n, t]}{\text{out\_deg}[n]}$   
**end for**  
**return**  $TER$

---

node capacity constraints. The selection process considers both topological constraints and temporal feasibility, ensuring that selected edges represent meaningful temporal interactions. For nodes in valid edges, we consider the temporally distant timestamp<sup>3</sup> (*SelectBest*). When the initial selection phase cannot fulfill the target edge budget, a recovery mechanism (*RecoverEdges*) is activated to identify additional valid edge-timestamp pairs while preserving the temporal consistency of the network. This approach ensures both structural balance through capacity constraints and temporal coherence through window-based selection, making it particularly suitable for analyzing dynamic interaction patterns in temporal networks.

---

#### Algorithm 2: Timestamp Selector (Main Procedure)

---

- 1: **Input:** Graph  $G(V, E)$ , Node capacity  $C$ , Time window  $W$
- 2: **Output:** Edge-timestamp pairs  $M$
- 3:  $M \leftarrow \emptyset$  {Final edge-timestamp assignments}
- 4:  $B \leftarrow \text{COMPUTEBUDGET}(C)$  {Total number of insertions}
- 5: Sort nodes  $v \in V$  by structural priority
- 6: **for** all  $v \in V$  **do**
- 7:     **while** HASCAPACITY( $v$ ) **and**  $|M| < B$  **do**
- 8:          $E_c \leftarrow \text{GETVALIDEDGES}(v, W)$
- 9:         **if**  $E_c \neq \emptyset$  **then**
- 10:              $(e, t) \leftarrow \text{SELECTBEST}(E_c)$
- 11:              $M \leftarrow M \cup \{(e, t)\}$
- 12:             UPDATECAPACITY( $e$ )
- 13:         **end if**
- 14:     **end while**
- 15: **end for**
- 16: **if**  $|M| < B$  **then**
- 17:      $M \leftarrow M \cup \text{RECOVEREDGES}(B - |M|)$
- 18: **end if**
- 19: **return**  $M$

---

#### Helper Routines.

- COMPUTEBUDGET( $C$ ): Calculates the total number of

<sup>3</sup>We select a temporally distant timestamp for these nodes because it will be the least likely interaction for these nodes.

edges to be added, based on node-level capacity  $C$ .

- HASCAPACITY( $v$ ): Returns true if node  $v$  has not exceeded its insertion/deletion budget.
- GETVALIDEDGES( $v, W$ ): Returns a set of candidate edges involving  $v$  within the time window  $W$ , filtered by feasibility constraints (e.g., C3/C4).
- SELECTBEST( $E_c$ ): Chooses the most adversarially valuable edge–timestamp pair from the candidate set, based on sparsification metrics.
- UPDATECAPACITY( $e$ ): Updates the capacity counters for nodes in edge  $e$ .
- RECOVEREDGES( $k$ ): Backup strategy to find additional  $k$  edge–timestamp pairs when primary routine falls short (e.g., due to constraint violations).

**Why can’t other negative sampling algorithms be used instead of LORETTA’s:** Regarding negative sampling, random sampling (selecting edges arbitrarily out of temporal order) and algorithms like Havel-Hakimi (Kleitman and Wang 1973) are insufficient for the adversarial setting as they fail to account for stealth and temporal constraints. Random sampling lacks strategic impact, while Havel-Hakimi focuses on degree preservation without addressing temporal dynamics or unnoticeability requirements. TERA utilizes a novel backtracking algorithm to replace critical edges with carefully selected negative samples that preserve graph density and temporal structure, ensuring the perturbations remain subtle yet impactful. This combination of temporal PageRank-based edge selection and backtracking ensures an effective and generalizable adversarial attack framework tailored for temporal graph networks.

### A.15 Why is REM Attack on Wikipedia so Effective?

In our experiments, we observed that applying sparsification *alone* on the Wikipedia dataset consistently outperforms combined strategies that include both sparsification and negative sampling (NS). We hypothesize that this is due to the unique structural and semantic properties of the Wikipedia graph.

**Semantic Modularity & Implicit Regularization**  
 Wikipedia exhibits a highly modular interaction structure, where users predominantly contribute to domain-specific clusters (e.g., nuclear physics vs. fashion). Sparsification selectively removes semantically meaningful edges within these modules, directly impairing the model’s ability to capture local coherence. In contrast, negative sampling often introduces edges between semantically distant nodes, implicitly injecting noise that violates this modularity. While such perturbations are adversarial in nature, they may inadvertently regularize the TGNN by encouraging cross-domain generalization, smoothing latent representations across otherwise disjoint clusters.

On semantically modular graphs like Wikipedia, sparsification alone is highly effective as it removes high-signal, intra-cluster edges critical to the model’s inference. In contrast, adding semantically inconsistent edges via NS can counteract

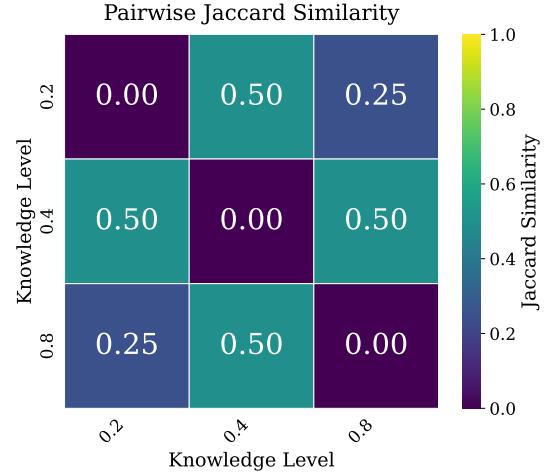


Figure 6: Pairwise Jaccard similarity between edge sets sparsified under different adversary knowledge levels ( $k$ ). As knowledge increases, the edge sets diverge in a structured and deterministic manner.

this effect by implicitly encouraging broader generalization. This interplay likely explains why REM attacks on Wikipedia are more damaging than their LORETTA counterparts.

### A.16 Ablation on Knowledge-Conditioned Sparsification

To analyze the behavior of our sparsification method under varying levels of adversary knowledge, we perform a controlled ablation by sparsifying the same dataset using progressively increasing knowledge thresholds:  $k \in \{0.2, 0.4, 0.8\}$ . For each threshold, we extract the resulting edge set and compute pairwise Jaccard similarity between all combinations. The resulting similarity matrix is visualized in Figure 6.

Our method exhibits three key properties that differentiate it from baseline random removal. First, it is *deterministic*—the exact same edge set is produced for a given knowledge level across multiple runs, indicating no inherent randomness in the sparsification pipeline. This ensures reproducibility and enhances interpretability of downstream behavior. Second, the method displays *monotonicity* in its edge removal: as the adversary’s knowledge increases from  $k_1$  to  $k_2$  with  $k_1 < k_2$ , the edge set removed under  $k_1$  is always a strict subset of the set removed under  $k_2$ . This is validated both by the decreasing Jaccard similarities and the subset relationships observed in our overlap analysis. Third, we observe a consistent trend of *adversarial alignment*: higher knowledge levels lead to more aggressive removal of structurally important edges, suggesting that the method adapts meaningfully to the threat model. In contrast, randomly sparsified graphs show stable performance across varying knowledge levels, which we attribute to a non-monotonic relationship between edge removal and attack effectiveness. In such cases, only targeted perturbations cause significant degradation, while excessive or uninformed removals disrupt the attacker’s influence. This implies a “sweet spot” of adversary knowl-

edge, where the attack is both precise and effective—too little misses key substructures, and too much reduces the coherence of the perturbation. Collectively, these observations highlight that our method produces more structured and predictable defense responses than random strategies.

### A.17 Does Cosine and T-shield Filtering remove true edges?

Table 9 shows the results of the average % of adversarial edges removed during the filtering-based adversarial defense methods for Wikipedia and UCI on TGN when perturbed using LORETTA at  $p = 0.3$ . We can clearly observe that in most of the cases the % of adversarial edges being removed is around 30% which means many true edges are being removed which explains why after applying filtering based adversarial defense methods the performance is even worse than after LORETTA.

Attack	Dataset	Cosine	T-shield-F	T-shield
Degree	Wikipedia	32.25	33.21	32.50
	UCI	35.08	42.09	41.72
PageRank	Wikipedia	32.09	33.68	33.51
	UCI	32.48	39.07	39.48
Cosine	Wikipedia	35.92	30.95	30.85
	UCI	29.46	29.18	28.77
Jaccard	Wikipedia	38.44	30.70	30.77
	UCI	30.72	30.86	30.98
TER	Wikipedia	38.11	32.66	32.62
	UCI	34.38	34.15	33.66

Table 9: Average % of adversarial edges removed during filtering-based adversarial defense methods.

### A.18 Threshold Free summary results of LORETTA against SotA Anomaly Detection methods

Attack	Anomaly Detection Methods (AUPRC)			
	MIDAS	F-FADE	AnoEdge-L	AnoEdge-G
<b>Wikiedia Dataset</b>				
Degree	0.44	0.43	0.48	0.45
PageRank	0.45	0.44	0.49	0.46
Cosine	0.40	0.37	0.45	0.58
Jaccard	0.32	0.35	0.43	0.54
TER	0.50	0.34	0.40	0.34
<b>UCI Dataset</b>				
Degree	0.37	0.55	0.24	0.31
PageRank	0.46	0.56	0.26	0.36
Cosine	0.44	0.28	0.46	0.80
Jaccard	0.48	0.24	0.44	0.83
TER	0.72	0.19	0.43	0.56

Table 10: Anomaly-class AUPRC (area under the precision-recall curve) of anomaly detection methods after perturbation using LORETTA.

Table 10 shows the results of AUPRC (area under precision recall curve) score of anomaly detection methods after per-

turbation using LORETTA at  $p = 0.3$ . We can see that the AUPRC is around 0.5 for most of the cases except for 3 places in the UCI dataset. However, it must be noted that this score represents all the thresholds and it is impossible to determine which threshold to use in real-world making detecting adversarial edges even more challenging. After this, we can confidently say that LORETTA is undetectable using SotA anomaly detection methods for edge streams.

### A.19 Statistical Confidence Results

For each *Model–Dataset* pair, we compare every LoReTTA variant against all available baselines using a common, scale-free degradation metric in Table 11. Given the clean validation score  $\text{Clean}_{m,d}$  for model  $m$  on dataset  $d$  and the post-attack score  $\text{Score}_{m,d}$ , we define the percentage drop

$$\text{Drop}\% = \frac{\text{Clean}_{m,d} - \text{Score}_{m,d}}{\text{Clean}_{m,d}} \times 100,$$

so that larger values indicate stronger attacks (greater degradation). A negative value (rare) means the attack improved the score relative to clean.

**Pairwise comparisons and outcomes.** For a fixed *Model–Dataset* pair, let  $\mathcal{B}$  denote the set of baselines with valid scores. For each LoReTTA variant  $v$  and each  $b \in \mathcal{B}$  we compare  $\text{Drop}\%_v$  to  $\text{Drop}\%_b$ . We record a *win* if  $\text{Drop}\%_v > \text{Drop}\%_b$ , a *loss* if  $\text{Drop}\%_v < \text{Drop}\%_b$ , and a *tie* otherwise. The table lists the counts of wins/losses/ties and the non-tied win rate.

**Sign test over baselines.** Because per-seed measurements are unavailable, we assess dominance using a one-sided exact binomial (sign) test across baselines:

$$H_0 : p \leq \frac{1}{2} \quad \text{vs.} \quad H_1 : p > \frac{1}{2},$$

where  $p$  is the probability that the LoReTTA variant wins against a randomly chosen baseline in  $\mathcal{B}$  (ties are excluded). Given  $W$  wins and  $L$  losses, the  $p$ -value is computed under  $\text{Binomial}(W+L, \frac{1}{2})$ . We report this  $p$ -value per row and mark significance with stars: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . These stars summarize statistical significance (not effect size).

**Interpretation and scope.** Rows with \*\*/\*\*\* indicate that, for the given *Model–Dataset*, the LoReTTA variant beats a clear majority of baselines with high confidence. Rows without stars mean the evidence is insufficient to claim dominance at the 0.05 level. All comparisons use the same  $\text{Drop}\%$  metric, making results comparable across datasets with different absolute scales.

Table 11: Sign-test results for LoReTTA variants vs. all baselines (higher Drop% is stronger). One-sided exact binomial test across baselines per Model–Dataset; ties excluded.

Model	Dataset	Variant	Drop%	Wins	Losses	Ties	Non-ties	Win rate	p-value	Sig.
DySAT	Enron	Degree	57.20	11	0	0	11	1.000	4.88e-04	***
DySAT	Enron	LoReTTA – Cosine	30.04	11	0	0	11	1.000	4.88e-04	***
DySAT	Enron	LoReTTA – Jaccard	35.31	11	0	0	11	1.000	4.88e-04	***
DySAT	Enron	LoReTTA – TER	24.12	11	0	0	11	1.000	4.88e-04	***
DySAT	Enron	Pagerank	54.07	11	0	0	11	1.000	4.88e-04	***
DySAT	UCI	Degree	72.44	11	0	0	11	1.000	4.88e-04	***
DySAT	UCI	LoReTTA – Cosine	66.89	11	0	0	11	1.000	4.88e-04	***
DySAT	UCI	LoReTTA – Jaccard	68.56	11	0	0	11	1.000	4.88e-04	***
DySAT	UCI	LoReTTA – TER	73.38	11	0	0	11	1.000	4.88e-04	***
DySAT	UCI	Pagerank	72.51	11	0	0	11	1.000	4.88e-04	***
DySAT	Wikipedia	Degree	22.25	5	4	0	9	0.556	5.00e-01	
DySAT	Wikipedia	LoReTTA – Cosine	27.34	5	4	0	9	0.556	5.00e-01	
DySAT	Wikipedia	LoReTTA – Jaccard	27.61	5	4	0	9	0.556	5.00e-01	
DySAT	Wikipedia	LoReTTA – TER	20.60	5	4	0	9	0.556	5.00e-01	
DySAT	Wikipedia	Pagerank	19.35	5	4	0	9	0.556	5.00e-01	
JODIE	Enron	Degree	56.59	11	0	0	11	1.000	4.88e-04	***
JODIE	Enron	LoReTTA – Cosine	22.59	11	0	0	11	1.000	4.88e-04	***
JODIE	Enron	LoReTTA – Jaccard	28.91	11	0	0	11	1.000	4.88e-04	***
JODIE	Enron	LoReTTA – TER	22.86	11	0	0	11	1.000	4.88e-04	***
JODIE	Enron	Pagerank	50.82	11	0	0	11	1.000	4.88e-04	***
JODIE	MOOC	Degree	69.36	9	0	0	9	1.000	1.95e-03	**
JODIE	MOOC	LoReTTA – Cosine	68.50	9	0	0	9	1.000	1.95e-03	**
JODIE	MOOC	LoReTTA – Jaccard	70.80	9	0	0	9	1.000	1.95e-03	**
JODIE	MOOC	LoReTTA – TER	61.42	3	6	0	9	0.333	9.10e-01	
JODIE	MOOC	Pagerank	67.94	9	0	0	9	1.000	1.95e-03	**
JODIE	UCI	Degree	59.96	9	2	0	11	0.818	3.27e-02	*
JODIE	UCI	LoReTTA – Cosine	61.65	9	2	0	11	0.818	3.27e-02	*
JODIE	UCI	LoReTTA – Jaccard	63.21	11	0	0	11	1.000	4.88e-04	***
JODIE	UCI	LoReTTA – TER	63.01	11	0	0	11	1.000	4.88e-04	***
JODIE	UCI	Pagerank	63.86	11	0	0	11	1.000	4.88e-04	***
JODIE	Wikipedia	Degree	50.40	6	3	0	9	0.667	2.54e-01	
JODIE	Wikipedia	LoReTTA – Cosine	55.46	7	2	0	9	0.778	8.98e-02	
JODIE	Wikipedia	LoReTTA – Jaccard	54.57	7	2	0	9	0.778	8.98e-02	
JODIE	Wikipedia	LoReTTA – TER	46.42	6	3	0	9	0.667	2.54e-01	
JODIE	Wikipedia	Pagerank	51.50	6	3	0	9	0.667	2.54e-01	
TGAT	Enron	Degree	58.22	11	0	0	11	1.000	4.88e-04	***
TGAT	Enron	LoReTTA – Cosine	48.18	11	0	0	11	1.000	4.88e-04	***
TGAT	Enron	LoReTTA – Jaccard	48.33	11	0	0	11	1.000	4.88e-04	***
TGAT	Enron	LoReTTA – TER	21.67	11	0	0	11	1.000	4.88e-04	***
TGAT	Enron	Pagerank	57.95	11	0	0	11	1.000	4.88e-04	***
TGAT	UCI	Degree	34.97	11	0	0	11	1.000	4.88e-04	***
TGAT	UCI	LoReTTA – Cosine	28.65	10	1	0	11	0.909	5.86e-03	**
TGAT	UCI	LoReTTA – Jaccard	31.10	10	1	0	11	0.909	5.86e-03	**
TGAT	UCI	LoReTTA – TER	33.99	11	0	0	11	1.000	4.88e-04	***
TGAT	UCI	Pagerank	32.21	10	1	0	11	0.909	5.86e-03	**
TGAT	Wikipedia	Degree	4.42	4	5	0	9	0.444	7.46e-01	
TGAT	Wikipedia	LoReTTA – Cosine	16.61	4	5	0	9	0.444	7.46e-01	
TGAT	Wikipedia	LoReTTA – Jaccard	13.60	4	5	0	9	0.444	7.46e-01	
TGAT	Wikipedia	LoReTTA – TER	-7.82	1	8	0	9	0.111	9.98e-01	
TGAT	Wikipedia	Pagerank	1.58	2	7	0	9	0.222	9.80e-01	
TGN	Enron	Degree	55.88	11	0	0	11	1.000	4.88e-04	***
TGN	Enron	LoReTTA – Cosine	27.24	11	0	0	11	1.000	4.88e-04	***
TGN	Enron	LoReTTA – Jaccard	34.48	11	0	0	11	1.000	4.88e-04	***
TGN	Enron	LoReTTA – TER	31.90	11	0	0	11	1.000	4.88e-04	***
TGN	Enron	Pagerank	51.97	11	0	0	11	1.000	4.88e-04	***
TGN	MOOC	Degree	67.46	9	0	0	9	1.000	1.95e-03	**
TGN	MOOC	LoReTTA – Cosine	62.38	9	0	0	9	1.000	1.95e-03	**
TGN	MOOC	LoReTTA – Jaccard	62.93	9	0	0	9	1.000	1.95e-03	**

*Continued on next page*

Model	Dataset	Variant	Drop%	Wins	Losses	Ties	Non-ties	Win rate	p-value	Sig.
TGN	MOOC	LoReTTA – TER	50.32	8	1	0	9	0.889	1.95e-02	*
TGN	MOOC	Pagerank	68.04	9	0	0	9	1.000	1.95e-03	**
TGN	UCI	Degree	60.38	11	0	0	11	1.000	4.88e-04	***
TGN	UCI	LoReTTA – Cosine	65.11	11	0	0	11	1.000	4.88e-04	***
TGN	UCI	LoReTTA – Jaccard	64.59	11	0	0	11	1.000	4.88e-04	***
TGN	UCI	LoReTTA – TER	65.02	11	0	0	11	1.000	4.88e-04	***
TGN	UCI	Pagerank	61.13	11	0	0	11	1.000	4.88e-04	***
TGN	Wikipedia	Degree	17.83	5	4	0	9	0.556	5.00e-01	
TGN	Wikipedia	LoReTTA – Cosine	32.86	6	3	0	9	0.667	2.54e-01	
TGN	Wikipedia	LoReTTA – Jaccard	31.45	6	3	0	9	0.667	2.54e-01	
TGN	Wikipedia	LoReTTA – TER	16.19	5	4	0	9	0.556	5.00e-01	
TGN	Wikipedia	Pagerank	17.38	5	4	0	9	0.556	5.00e-01	

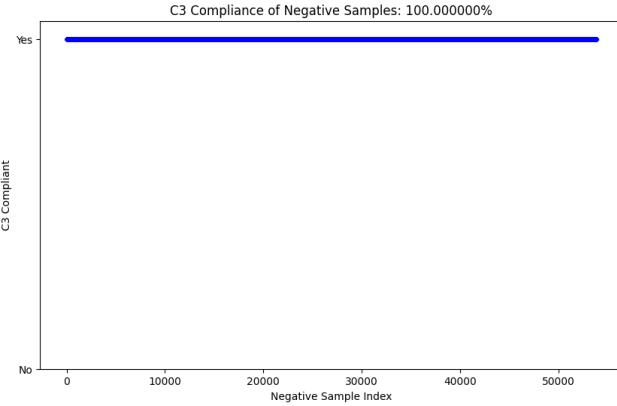


Figure 7: C3 compliance for the 400s dataset. All negative samples satisfy the C3 constraint, indicated by blue dots.

## A.20 Effect of Time Window on C3 and C4 Constraints

Figure 7 highlights that the 400s time window satisfies the C3 compliance condition. In contrast, the 1800s window fails to maintain this distribution (Figure 8).

However, the opposite trend appears in the C4 constraint. The 1800s window exhibits better degree preservation (Figure 9), whereas the 400s window shows notable deviation (Figure 10). This is likely due to sparsity in the 400s, where timestamps can have zero active nodes because of smaller time-window.

Our main experiments use the time-window, guided by Lee, Lee, and Shin (2024) on TGNN degradation.

## Ethical Considerations

This work presents a novel adversarial attack on TGNNs, aiming to identify vulnerabilities in temporal graph learning models. While adversarial attacks have the potential to be misused, our primary motivation is to improve the robustness of temporal graph-based models by identifying weaknesses that could be exploited in malicious scenarios. By understanding these vulnerabilities, we hope to enable the development of defense mechanisms and inspire more resilient temporal graph architectures.

This work could benefit areas such as network security, fraud detection, and dynamic system modeling by highlighting risks and promoting research on robust temporal graph methods. However, the methods described in this paper, if misapplied, could be used to undermine real-world systems relying on temporal graphs. We strongly advocate for ethical use of this research, emphasizing its role in advancing security and resilience in machine learning.

As with all adversarial research, this work underscores the importance of ethical responsibility in deploying machine learning models in safety-critical applications. Future research could focus on leveraging insights from this work to develop robust defenses against such attacks and to minimize the risks associated with deploying temporal graph models in sensitive environments.

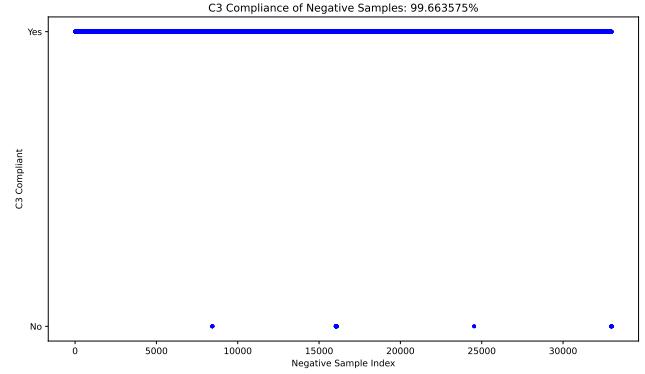


Figure 8: C3 compliance for the 1800s dataset. Blue dots at top indicate compliant samples; a small fraction at bottom BC violate the C3 constraint.

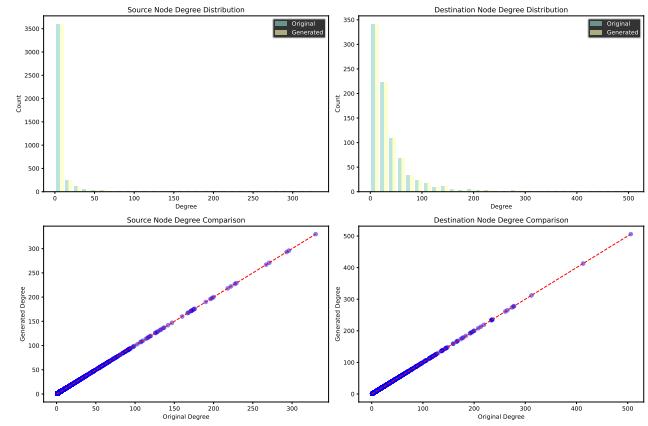


Figure 9: C4 compliance (degree preservation) for 1800s. Scatter alignment confirms statistical matching. Original vs regenerated degrees are identical.

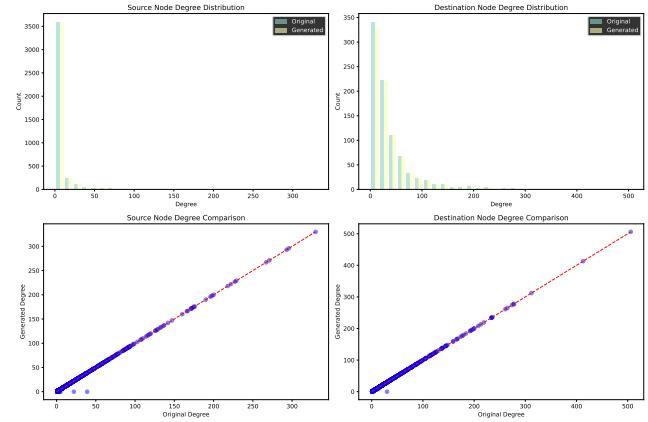


Figure 10: C4 compliance for the 400s dataset. While overall degree preservation holds, early timestamps show deviation due to low activity. Many nodes were yet to be created, reducing the likelihood of matching active degrees.