PROPAGATION IDENTITY PERTURBATIONS: A NEW ADVERSARIAL THREAT MODEL FOR ROBUST SOCIAL EVENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Social event forecasting models depend critically on who participates in the propagation process, which differs from conventional graph learning tasks. Existing robustness research in this area intensively studies perturbations on graph topology, node features, or textual content, but leaves propagation identities underexplored. We propose Propagation Identity Perturbation (PIP), a new adversarial threat model that formalizes attacks on participant identities in temporal diffusion trajectories while preserving realistic propagation behavior. We instantiate PIP as a gradient-guided discrete attack that searches behaviorally plausible identity substitutions within constrained candidate spaces. Moreover, we provide theoretical support for PIP in feasible search space reduction, transferability lower bound, monotonic ascent property, and guaranteed loss disruption bound. Extensive experiments on five representative social event forecasting architectures against five mainstream defense measures further expose the severity of this overlooked vulnerability in diffusion forecasting systems. These findings highlight an urgent need to develop identity-aware defense mechanisms to defend social event forecasting models against PIP.

1 INTRODUCTION

Robustness has become a central topic in graph representation learning (Shang et al., 2025). Existing adversarial attacks manipulate graph topology, node attributes, injected nodes, or textual inputs to degrade model performance. However, they mainly assumes that the adversary modifies the graph itself, which can be violated in social event forecasting (SEF). Predictions in SEF are generated from evolving propagation trajectories rather than static graph structures (Yuan et al., 2020; Sun et al., 2022; Cheng et al., 2023; Qiao et al., 2023; Wang et al., 2024). A propagation trajectory consists of a sequence of user participation events occurring over time, where each participant influences subsequent diffusion dynamics. Consequently, the natural attack surface is no longer graph topology or node features, but **the identities of propagation participants**. Currently, this attack surface has received little systematic investigation, which motivates us to constitute a new adversarial learning problem.

We formulate Propagation Identity Perturbation (PIP) as a new threat model for temporal propagation learning. Instead of modifying edges or node attributes, PIP perturbs participant identities in future propagation trajectories. Given a propagation sequence, the adversary replaces a small number of participant identities with alternative users while preserving behavioral plausibility constraints derived from temporal activity consistency, propagation semantics, and candidate legitimacy. This formulation introduces several new challenges: **1.** Propagation identities are **discrete combinatorial variables**, which makes continuous gradient perturbations inapplicable. **2.** Substitutions must remain **behaviorally plausible**. Arbitrary identity replacement often produces impossible propagation trajectories that violate realistic user activity patterns. **3.** Perturbations should **transfer across forecasting architectures**, which requires model-agnostic attack generation rather than victim-specific optimization. These characteristics distinguish PIP from existing graph adversarial settings and require new optimization strategies.

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

Moreover, we instantiate PIP as a discrete gradient-guided attack that operates over propagation identity space. Specifically, it computes identity importance from forecasting gradients while restricting candidate substitutions to a behaviorally plausible attack space. The candidate space is constructed from propagation-consistent users that satisfy temporal activity and diffusion legitimacy constraints, which enforces generated perturbations to remain realistic. Gradient-guided discrete search then selects identity substitutions that maximize forecasting disruption under a fixed perturbation budget. More importantly, PIP generates a single transferable attack population independent of target (victim) architectures. This design enables consistent robustness evaluation across heterogeneous forecasting models without architecture-specific optimization.

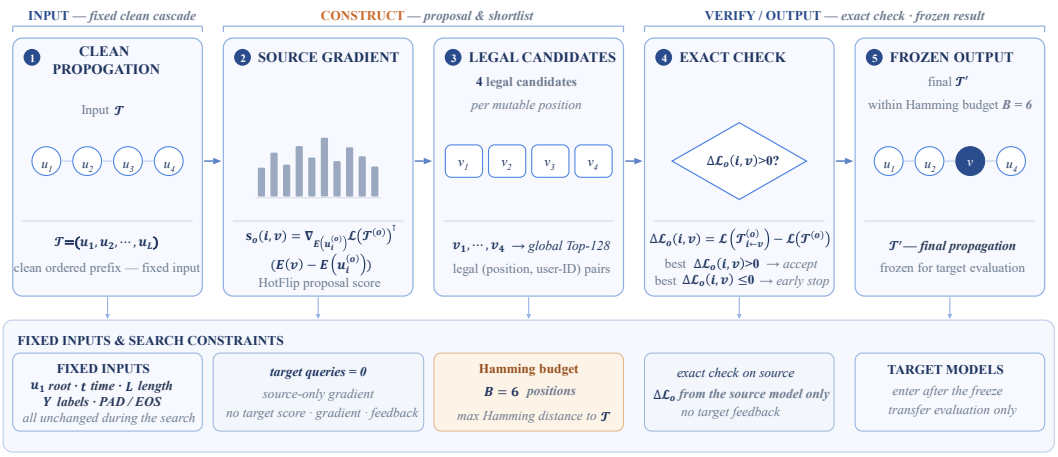

Figure 1: Threat model of the proposed PIP. It uses a gradient-guided discrete search, generates transferable attacks that significantly disrupt prediction accuracy across diverse forecasting architectures, and exposes vulnerabilities unaddressed by standard graph defenses.

The extensive experiments reveal a new phenomenon that existing defenses show significantly different robustness profiles across forecasting architectures, despite being designed for similar graph learning objectives. Through detailed analysis, we find that vulnerability is primarily determined by propagation-stage sensitivity rather than conventional graph structural sensitivity, since the dependence of forecasting predictions on identities appear at different stages of diffusion. This finding suggests that robustness in temporal diffusion forecasting follows a fundamentally different mechanism from robustness in static graph prediction and motivates new defense designs tailored to propagation dynamics. Our contributions are summarized as follows:

1. We introduce PIP, the first adversarial threat model targeting participant identities in temporal propagation trajectories. PIP formalizes robustness for social event forecasting from the perspective of propagation participants rather than graph topology or node features.

2. We propose a behaviorally constrained discrete attack that instantiates PIP through gradient-guided identity substitution in plausible propagation candidate spaces. The attack generates transferable adversarial propagation trajectories while preserving realistic propagation behavior.

3. We establish theoretical support for PIP across four perspectives: feasible search space reduction, transferability lower bound across models, monotonic ascent property, and guaranteed loss disruption bound.

2 Related Work

Graph Adversarial Robustness. Adversarial robustness in graph representation learning is commonly formulated as constrained optimization over graph topology or node attributes (Zügner et al., 2018; Zügner & Günnemann, 2019). Let a graph prediction model be $f_\theta : G = (V, E, X) \to y$. Then existing attacks optimize

$$\max_{\Delta E, \Delta X} \mathcal{L}\left(f_\theta(V, E + \Delta E, X + \Delta X), y\right), \tag{1}$$

subject to perturbation budgets on edges or node features. This formulation covers topology attacks, feature attacks, node injection, and poisoning attacks developed for graph neural networks and graph

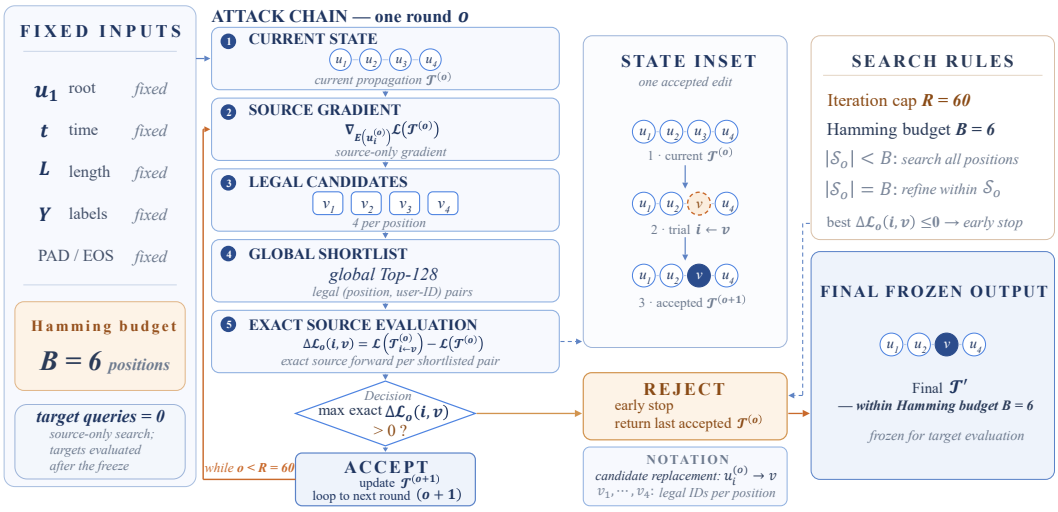

Figure 2: Attack chain of the proposed PIP. It iteratively generates candidate user-ID replacements with source-only gradients, shortlists the top legal position-user pairs, and accepts the update that maximizes exact source loss gains within a Hamming budget of B.

transformers (Dai et al., 2018; Sun et al., 2020; Geisler et al., 2021). Although these methods differ in optimization strategy, they all perturb the graph representation while **keeping propagation identities fixed**. Our work considers a different perturbation space, where the graph remains unchanged and the adversary instead perturbs propagation participants.

Social Event Forecasting. Social event forecasting predicts future diffusion dynamics from partially observed propagation trajectories (Li et al., 2017; Cao et al., 2017). Given a propagation prefix $\mathcal{T}_{1:k} = (u_i, t_i)_{i=1}^{k}$, forecasting models learn $f_\theta : (G, \mathcal{T}_{1:k}) \rightarrow y$, where user identities u_i are part of the predictive input (Yuan et al., 2020; Sun et al., 2022; Cheng et al., 2023; Qiao et al., 2023; Wang et al., 2024). However, existing robustness studies inherit perturbation models from generic graph learning or inject synthetic feature noise, most of which assume the adversarial propagation prefix $\mathcal{T}'$ to be identical to $\mathcal{T}$. In contrast, we study adversarial perturbations over the identity sequence itself. In the rest of this paper, we use $\mathcal{T}'$ to denote the **perturbed version** of $\mathcal{T}$.

Discrete Propagation Identity Attacks. Discrete adversarial attacks optimize perturbations over categorical variables via gradient-guided substitution or combinatorial search (Ebrahimi et al., 2018; Jin et al., 2020; Zou et al., 2023). A first-order approximation evaluates replacing a discrete variable by $\Delta\mathcal{L} \approx \nabla_{e_i}\mathcal{L}^\top (e(c) - e_i)$, where e_i denotes the embedding of the current discrete variable (Ebrahimi et al., 2018). Our setting differs because candidate substitutions are constrained by **behavioral plausibility** rather than vocabulary validity. For each propagation participant, $u_i' \in \mathcal{C}(u_i, t_i, G)$, where $\mathcal{C}$ enforces temporal activity consistency, propagation legitimacy, and neighborhood compatibility. Consequently, PIP optimizes over a constrained propagation trajectory rather than an unrestricted discrete sequence.

Transfer Robustness for Propagation Forecasting. Transfer attacks evaluate perturbations generated from one source model on unseen target models (Liu et al., 2017; Dong et al., 2018; Demontis et al., 2019). Existing graph robustness benchmarks typically generate perturbations independently for each target architecture, making transfer comparisons inconsistent. We instead construct a fixed adversarial population $\mathcal{A} = \{\mathcal{T}_1', \ldots, \mathcal{T}_N'\}$ shared across all target models. Every target model is evaluated on identical adversarial trajectories

$$R(f) = \mathbb{E}_{\mathcal{T}' \sim \mathcal{A}}\big[M(f, \mathcal{T}')\big], \tag{2}$$

which isolates architectural robustness from attack generation. This standardized transfer protocol complements the proposed propagation identity threat model and enables reproducible robustness evaluation across heterogeneous forecasting architectures.

3 METHODOLOGY

3.1 PROBLEM FORMULATION: PROPAGATION IDENTITY PERTURBATION (PIP)

We consider the task of SEF, where the objective is to predict the future evolution of an online information cascade from a partially observed propagation trajectory. Let $\mathcal{T} = (u_i, t_i)_{i=1}^{L}$ be an observed propagation trajectory consisting of an ordered sequence of user participation events, where $u_i \in \mathcal{U}$ and t_i denote the participating user ($\mathcal{U}$ being the universe of users) and the participation timestamp, respectively. A forecasting model receives a prefix $\mathcal{T}_{1:k} = (u_i, t_i)_{i=1}^{k}$ and predicts future propagation outcomes $y = f_\theta(G, \mathcal{T}_{1:k})$, where G denotes the underlying social graph and f_θ is a forecasting architecture (graph neural network, diffusion model, heterogeneous graph model, etc.). Unlike conventional graph prediction tasks, the prediction depends jointly on $(G, \mathcal{T}_{1:k})$, where $\mathcal{T}_{1:k}$ contains **identity information** describing *who* propagated the event.

We introduce PIP, a new adversarial threat model that perturbs propagation participants rather than graph topology. A perturbation replaces a subset of identities $\mathcal{S} \subseteq \{1, \ldots, k\}$ with alternative users $u_i \to u_i', i \in \mathcal{S}$. The perturbed trajectory becomes $\mathcal{T}' = \{(\tilde{u}_i, t_i)\}_{i=1}^{k}$, where $\tilde{u}_i = \begin{cases} u_i', & \text{if } i \in \mathcal{S}, \\ u_i, & \text{otherwise.} \end{cases}$. Importantly, timestamps and graph structure remain unchanged.

3.2 THREAT MODEL AND OPTIMIZATION OBJECTIVE

The adversary satisfies three constraints:
1. Budget constraint: Only a limited number of propagation identities may be modified, $|\mathcal{S}| \leqslant B$.
2. Discrete identity constraint: Each replacement must belong to the user universe, $u_i' \in \mathcal{U}$.
3. Behavioral plausibility constraint: Every replacement must remain behaviorally consistent with the original propagation process, $u_i' \in \mathcal{C}(u_i, t_i, G)$, where $\mathcal{C}$ denotes a behaviorally plausible candidate set defined later.
The adversary seeks perturbations that maximize forecasting loss:

$$\max_{\mathcal{T}' \in \Pi_B(\mathcal{T})} \mathcal{L}\left(f_\theta(G, \mathcal{T}'), y\right), \tag{3}$$

where $\Pi_B(\mathcal{T})$ denotes all feasible propagation identity perturbations satisfying the above constraints. This formulation differs fundamentally from feature perturbations or topology attacks because the optimization is performed over a discrete combinatorial identity space (see Table 1).

Table 1: PIP is fundamentally different from graph attacks.

Existing Graph Robustness	**Propagation Identity Perturbation (PIP)**
Perturb node features.	Perturb propagation participants.
Flip or inject graph edges.	Replace users while preserving graph topology.
Continuous feature optimization.	Discrete identity optimization.
Topology plausibility rarely modeled.	Behavioral plausibility is mandatory.

The optimization problem (3) is computationally challenging for three reasons:
1. Discrete combinatorial search: The feasible perturbation space grows exponentially, $|\Pi_B(\mathcal{T})| = \sum_{b=0}^{B} \binom{k}{b} (|\mathcal{U}| - 1)^b$. Even for modest propagation lengths, exhaustive search is impossible.
2. Behavioral validity: Replacing identities arbitrarily often produces impossible propagation trajectories, such as inactive users participating at unrealistic timestamps. Therefore perturbations must satisfy $P(\mathcal{T}') \geqslant \tau$, where P measures propagation plausibility.
3. Transferability: The attack should **generalize across different target forecasting architectures**: $f_{\theta_1}, f_{\theta_2}, \ldots$, without re-optimizing perturbations for each target model.
These challenges motivate the construction of a constrained attack space.

3.3 Behaviorally Plausible Attack Space

The key insight behind PIP is that realistic propagation attacks should remain indistinguishable from naturally occurring participant substitutions. For every propagation position i, we define a candidate set $\mathcal{C}_i = \mathcal{C}(u_i, t_i, G)$. Rather than searching over the entire user universe, the adversary searches only within $\mathcal{C}_i$. Each candidate satisfies three behavioral constraints:

1. Temporal consistency. Candidate users must be active during the same temporal window:

$$\mathcal{C}_i^{(t)} = \{v \in \mathcal{U} : \exists\, \iota \in \mathrm{Act}(v) :\ |\iota - t_i| \leqslant \rho\}, \tag{4}$$

where $\mathrm{Act}(v)$ is the set of activity timestamps of v on the training split, and ρ is the temporal tolerance.

2. Propagation consistency. Candidates show similar propagation participation characteristics:

$$\mathcal{C}_i^{(p)} = \{v \in \mathcal{U} : v \in \mathrm{Role}\big(\mathfrak{b}(i)\big)\}, \tag{5}$$

where $\mathrm{Role}(\mathfrak{b})$ is the set of users **observed in relative position bin** $\mathfrak{b}$ on the training split, where relative position bins are defined by $\mathfrak{b}(i) = \lfloor 5i/(L-1) \rfloor$ (five equally spaced bins) and L is the propagation length.

3. Neighborhood consistency. Candidates occupy comparable structural neighborhoods:

$$\mathcal{C}_i^{(n)} = \{v \in \mathcal{U} : d_G(u_{i-1}, v) \leqslant r\}, \tag{6}$$

where $d_G(\cdot,\cdot)$ is the **hop distance** on the social graph G. In practice, it is approximated by statistics collected on the training split. $d_G(u_{i-1}, v) \leqslant r$ means that v lies in the r-hop neighborhood of u_{i-1}, which is the r-fold iteration of $\mathrm{Step}(\cdot)$:

$$\mathrm{Step}(u_{i-1}) = \mathscr{S}(u_{i-1}) \cup \mathcal{N}_G(u_{i-1}),\ \mathscr{S}(u) = \{v : (u, v) \text{ appear consecutively in a training cascade}\}. \tag{7}$$

Therefore, the one-hop neighborhood combines training transition successors and social-graph neighbors $\mathcal{N}_G$.

This construction guarantees that perturbations remain compatible with realistic diffusion behavior. Considering all three behavioral constraints, the final candidate set at position i becomes

$$\mathcal{C}_i = \left(\bigcap_{m \in \{t,p,n\}} \mathcal{C}_i^{(m)} \right) \setminus \big(\{u_j\}_{j=1}^{k} \cup \{u_i\} \big), \tag{8}$$

which removes users already present in the propagation and u_i itself to prevent duplicate identities. Offline preprocessing constructs all candidate sets. In the perspective of algorithm, it computes behavioral statistics for every user, builds nearest-neighbor candidate indices, and retains candidates that satisfy the plausibility constrains. This preprocessing is independent of target models and performed only once. The resulting candidate sets significantly reduce the combinatorial search space while preserving realistic perturbations, which will be explained in Theorem 1.

3.4 Gradient-guided Propagation Identity Attack

We now derive an efficient optimizer for the PIP objective. Each propagation identity is represented by an embedding $e_i = E(u_i)$, where E denotes the shared user embedding matrix. The forecasting loss and its gradient with respect to (w.r.t.) propagation embeddings are:

$$\mathcal{L}\left(f_\theta(G, \mathcal{T}), y\right), \quad g_i = \nabla_{e_i} \mathcal{L}. \tag{9}$$

The gradient measures the local sensitivity of forecasting performance to each propagation participant. Replacing u_i with candidate v changes the embedding by $(E(v) - E(u_i))$. Using first-order approximation,

$$\Delta\mathcal{L} \approx g_i^\top \left(E(v) - E(u_i)\right). \tag{10}$$

The optimal candidate solves

$$v^* = \arg\max_{v \in \mathcal{C}_i} g_i^\top \left(E(v) - E(u_i)\right), \tag{11}$$

Algorithm 1 Gradient-guided propagation identity attack

Require: Clean propagation trajectory $\mathcal{T}$, labels Y, source model f_θ, candidate sets $\{\mathcal{C}_i\}_{i=1}^{k}$, modification budget $B = 6$, proposal steps $R = 60$.

$\mathcal{T}^{(0)} \leftarrow \mathcal{T}$.
for $o = 0, \ldots, R-1$ **do**
  $\mathcal{S}_o \leftarrow \{i : \mathcal{T}_i^{(o)} \neq \mathcal{T}_i\}$.
  $\mathcal{I}_o \leftarrow \mathcal{I}_{\text{mutable}}$ if $|\mathcal{S}_o| < B$, else $\mathcal{I}_o \leftarrow \mathcal{S}_o$, where $\mathcal{I}_{\text{mutable}}$ denotes the set of all eligible position indices in the propagation sequence $\mathcal{T}_{1:k}$ that the adversary is allowed to modify.
  Compute source loss gradients at $\mathcal{T}^{(o)}$ for positions in $\mathcal{I}_o$.
  Retain four legal replacement users $v \in \mathcal{C}_i$ per position $i \in \mathcal{I}_o$ that maximizes first-order score $\mathfrak{s}_o(i, v)$.
  $\mathcal{P}_o \leftarrow$ global top 128 legal position–identity pairs (i, v) with $v \in \mathcal{C}_i$.
  if $\mathcal{P}_o = \varnothing$ **then**
    break.
  end if
  Evaluate $\Delta\mathcal{L}_o(i, v)$ exactly with f_θ for every $(i, v) \in \mathcal{P}_o$.
  if $\max_{(i,v)\in\mathcal{P}_o} \Delta\mathcal{L}_o(i, v) \leqslant 0$ **then**
    break.
  end if
  $\mathcal{T}^{(o+1)} \leftarrow \mathcal{T}^{(o)}$ with the best pair applied.
end for
return the final perturbed $\mathcal{T}'$ with timestamps, length, root, and labels unchanged.

which yields an efficient discrete optimizer over plausible candidates. This optimization generalizes the HotFlip attack (Ebrahimi et al., 2018) to the PIP threat model.

Let $\mathcal{T}^{(0)} = \mathcal{T}$, let $\mathcal{T}^{(o)}$ be the current propagation trajectory after o accepted updates, and let $\mathcal{L}(\mathcal{T}^{(o)})$ be the source clean-label cross-entropy. For a legal replacement $u_i^{(o)} \leftarrow v$, the first-order proposal score is

$$\mathfrak{s}_o(i, v) = \nabla_{E(u_i^{(o)})} \mathcal{L}(\mathcal{T}^{(o)})^\top (E(v) - E(u_i^{(o)})). \tag{12}$$

At each search step, the method retains four candidates per mutable position and then the global top 128 legal position–identity pairs. Each shortlisted pair is evaluated by a complete source-model forward pass:

$$\Delta\mathcal{L}_o(i, v) = \mathcal{L}(\mathcal{T}_{i\leftarrow v}^{(o)}) - \mathcal{L}(\mathcal{T}^{(o)}). \tag{13}$$

Only the pair with the largest strictly positive exact gain is accepted. Before the budget B is full, every mutable position is eligible. Once B positions differ from the clean cascade, proposals are restricted to those B positions, so they may be refined without expanding the Hamming support. Search terminates after R proposal steps or at the first step with no legal positive-gain candidate. Exact verification **remains entirely on the source model** and therefore does **not violate the zero-target-query protocol**. The entire attack algorithm is summarized in Algorithm 1.

Protocol guarantee. By construction, the returned propagation satisfies all the constraints and remains within Hamming budget B. Every accepted update strictly increases the source loss, and no target appears in proposal, verification, or stopping. The search loop uses at most $R = 60$ gradient computations and $60 \times 128 = 7680$ candidate forward evaluations per cascade.

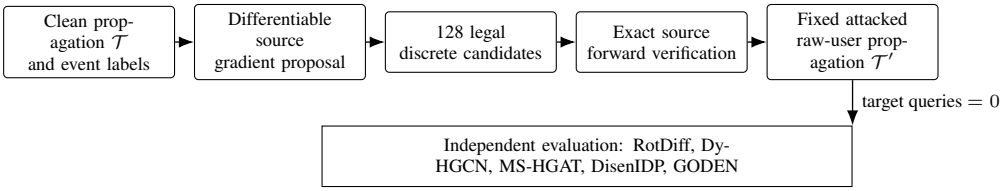

Figure 3: Separation between source-side attack generation and target-side evaluation. Only the differentiable source participates in proposal and exact verification. Every target receives the same fixed raw-user propagation and returns no feedback to the attack loop.

Proposal, verification, and query accounting. The gradient is a proposal mechanism rather than the acceptance criterion. Exact verification is necessary because replacing a complete identity can move far beyond the local linear approximation. At most 128 source forwards are performed per proposal step, and no score from target models enters the loop. Thus additional source computation can improve discrete selection without changing the zero-target-query property. The same generated raw-user file is reused for every target and every defense evaluation, which prevents target-specific sample selection.

4 Main Theoretical Results

We establish the theoretical foundations of PIP from four complementary perspectives: **(1)** feasible search space reduction, **(2)** transferability across forecasting models, **(3)** monotonic ascent for loss function, and **(4)** guaranteed loss disruption bound. Together, these results verify that PIP is not just a heuristic perturbation strategy, but a mathematically well-defined adversarial threat model with quantifiable properties. At the beginning, we introduce some preliminary assumptions.

Assumption 1 (Embedding Representation)**.** Let $h_\theta(\mathcal{T}) = \Phi_\theta(G, \mathcal{T})$ be the propagation representation learned by the forecasting encoder. The predictor is $f_\theta(G, \mathcal{T}) = g_\theta(h_\theta(\mathcal{T}))$.

Assumption 2 (Latent Representation Alignment)**.** The target model h_ϑ has close-enough latent representation alignment to the source model h_θ. There exists an orthogonal matrix W and $\epsilon > 0$,

$$\|h_\theta(\mathcal{T}) - W h_\vartheta(\mathcal{T})\|_2 \leqslant \epsilon, \quad \forall \mathcal{T}. \tag{14}$$

Assumption 3 (Latent Continuity and Loss Smoothness)**.** We assume latent continuity and loss smoothness in the target model as follows:

$$\begin{aligned} \forall \mathcal{T}, \mathcal{T}', \quad \|h_\vartheta(\mathcal{T}) - h_\vartheta(\mathcal{T}')\|_2 \leqslant & \mathscr{L}_h \sum_{i=1}^{k} \|E(u_i) - E(\tilde{u}_i)\|_2, \\ \|\nabla_h \mathcal{L}(h(\mathcal{T})) - \nabla_h \mathcal{L}(h(\mathcal{T}'))\|_2 \leqslant & \zeta \|h(\mathcal{T}) - h(\mathcal{T}')\|_2, \end{aligned} \tag{15}$$

where $\mathscr{L}_h$ and ζ denote the corresponding Lipschitz constants.

Theorem 1 (Feasible Search Space Reduction)**.** *Under the behavioral candidate constraint $\mathcal{C}$, suppose every propagation position admits at most m feasible replacement identities, where $m \leqslant (|\mathcal{U}| - 1)$. Let $\mathcal{S}$ be the perturbed positions. For fixed $|\mathcal{S}| = b$, the sizes of partial constrained and unconstrained search spaces satisfy*

$$\frac{N_b^{\mathcal{S}}}{N_b} \leqslant \left(\frac{m}{|\mathcal{U}| - 1}\right)^b, \quad 0 \leqslant b \leqslant B. \tag{16}$$

Summing over all valid perturbation budgets $b = |\mathcal{S}| \in \{0, 1, \ldots, B\}$ under candidate constraint $\mathcal{C}$, the entire constrained search space has size:

$$|\Pi_B^{\mathcal{C}}| = \sum_{b=0}^{B} N_b^{\mathcal{S}} \leqslant \sum_{b=0}^{B} \left(\frac{m}{|\mathcal{U}| - 1}\right)^b N_b \leqslant \sum_{b=0}^{B} N_b = |\Pi_B|. \tag{17}$$

Theorem 1 indicates that instead of searching over every user, PIP searches over a dramatically reduced combinatorial space. Taking the experimental setup in Appendix A.4 as an example, we have $|\mathcal{U}| = 12627$, $m = 6354$, and $B = 6$, thus $(\frac{m}{|\mathcal{U}|-1})^B = 0.0162$. Moreover, the median and mean numbers of feasible replacement identities w.r.t. different propagation positions are 3020 and 2917, respectively, which are **much smaller** than $m = 6354$. Thus the search space is **further reduced** from the above theoretical upper bound in **practical implementation**.

Theorem 2 (Transferability Lower Bound Across Models)**.** *Let*

$$\Delta h_\vartheta = h_\vartheta(\mathcal{T}') - h_\vartheta(\mathcal{T}), \quad \Delta h_\theta = h_\theta(\mathcal{T}') - h_\theta(\mathcal{T}). \tag{18}$$

Then under Assumptions 1 and 2,

$$\|\Delta h_\vartheta\|_2 \geqslant \|\Delta h_\theta\|_2 - 2\epsilon. \tag{19}$$

Theorem 2 indicates that if propagation encoders learn sufficiently aligned latent representations, perturbations that induce large representation displacement on one model can induce comparable displacement on another.

Theorem 3 (Monotonic Ascent Property of Algorithm 1)**.** *The generated propagation sequence* $\{\mathcal{T}^{(o)}\}$ *satisfies*

$$\mathcal{L}(\mathcal{T}^{(o+1)}) > \mathcal{L}(\mathcal{T}^{(o)}) \tag{20}$$

for every accepted iteration. Moreover, $\mathcal{L}(\mathcal{T}') \geqslant \mathcal{L}(\mathcal{T})$.

Theorem 3 establishes monotonic improvement of Algorithm 1, even without convexity.

Theorem 4 (Guaranteed Loss Disruption Bound)**.** *Let* $\mathcal{T}'$ *be obtained by replacing identities at indices* $\mathcal{S}$ *with* $|\mathcal{S}| \leqslant B$*. Define*

$$\Delta_{e_i} = E(u_i') - E(u_i), \Delta_H = h_\vartheta(\mathcal{T}') - h_\vartheta(\mathcal{T}). \tag{21}$$

Under Assumption 3,

$$\mathcal{L}(h_\vartheta(\mathcal{T}')) \geqslant \mathcal{L}(h_\vartheta(\mathcal{T})) + \nabla_h \mathcal{L}(h_\vartheta(\mathcal{T}))^\top \Delta_H - \frac{\zeta \mathscr{L}_h^2}{2} \left(\sum_{i \in \mathcal{S}} \|\Delta_{e_i}\|_2 \right)^2. \tag{22}$$

Theorem 4 reveals two key insights: **1.** The middle term $\nabla_h \mathcal{L}(h_\vartheta(\mathcal{T}))^\top \Delta_H$ represents the linear approximation of the loss increase. To maximize the loss, the attacker selects candidate identity substitutions whose embedding shifts align with the loss gradient. It validates the gradient-guided discrete replacement rule (11). **2.** The negative quadratic term $-\frac{\zeta \mathscr{L}_h^2}{2} \left(\sum_{i \in \mathcal{S}} \|\Delta_{e_i}\|_2 \right)^2$ acts as a penalty driven by the magnitude of identity embedding shifts and total modified positions $|\mathcal{S}|$. The attacker must constrain the Hamming budget $|\mathcal{S}| \leqslant B$ and select candidate replacements within plausible bounds to keep $\|\Delta_{e_i}\|_2$ small enough. The proofs of the above theorems are provided in Appendix A.1.

5 Experimental Results

The experimental setup extends the evaluation protocol of MILD (Shang et al., 2025) to our PIP threat model. While the Event-micro **Hits@10** in (48) serves as the main metric to evaluate the predicting performance of each case, we also calculate the following **Relative Hits@10 Drop (RHD)** to evaluate the attack disruption of PIP:

$$\text{RHD} = \frac{\text{Hits@10}_{\text{clean}} - \text{Hits@10}_{\text{undefended } or \text{ defended}}}{\text{Hits@10}_{\text{clean}}} \times 100\%. \tag{23}$$

The Twitter corpus distributed with the DisenIDP framework (Cheng et al., 2023) is used in the experiments. MILD (Shang et al., 2025) is used as the source model for PIP, while five state-of-the-art information diffusion prediction models RotDiff (Qiao et al., 2023), DyHGCN (Yuan et al., 2020), MS-HGAT (Sun et al., 2022), DisenIDP (Cheng et al., 2023), and GODEN (Wang et al., 2024) are used as targets. Besides, five representative defense methods: adversarial training (AT Madry et al. 2018), TRADES (Zhang et al., 2019), MART (Wang et al., 2020), SemRoDe-MMD (Formento et al., 2024), and ATFL (Wang et al., 2021), are used to assess the risk of PIP. The details of these target models and defense methods are provided in Appendices A.2 and A.3. Detailed experimental settings are provided in Appendix A.4. **Additional experimental results are provided in Appendix A.5, including ablation analysis and hyperparameter sensitivity analysis.**

5.1 Main Experimental Results

Table 2 reports the Hits@10 (%) and RHD (%) results across all evaluated settings under PIP. It shows the potency, cross-architecture transferability, and defense-evading capabilities of PIP as a new adversarial threat model.

Significant performance disruption. PIP achieves high attack efficacy across all evaluated target models without requiring any query to the target architectures during attack generation. On undefended target models, PIP causes an average macro Hits@10 drop from 34.02% to 25.29%, corresponding to an average RHD of 25.66%. The attack is particularly effective against higher-order hypergraph forecasting models and achieves peak RHDs of 30.91% on DisenIDP and 29.67%

Table 2: Main experimental results evaluating social event forecasting accuracy Hits@10 (%) and RHD (%) across five target models and five defense strategies under PIP attacks.

Defense	RotDiff	DyHGCN	MS-HGAT	DisenIDP	GODEN	Average
Clean	Hits@10 34.6442	Hits@10 33.1473	Hits@10 33.2234	Hits@10 34.5173	Hits@10 34.5685	Hits@10 34.0201
Undefended	Hits@10 27.0329 RHD **21.97**	Hits@10 25.3837 RHD **23.42**	Hits@10 23.3667 RHD **29.67**	Hits@10 23.8488 RHD **30.91**	Hits@10 26.8274 RHD **22.39**	Hits@10 25.2919 RHD **25.66**
AT	Hits@10 27.3119 RHD **21.16**	Hits@10 26.2083 RHD **20.93**	Hits@10 24.3055 RHD **26.84**	Hits@10 25.7136 RHD **25.51**	Hits@10 27.8553 RHD **19.42**	Hits@10 26.2789 RHD **22.75**
TRADES	Hits@10 27.0582 RHD **21.90**	Hits@10 25.1681 RHD **24.07**	Hits@10 23.3160 RHD **29.82**	Hits@10 24.9017 RHD **27.86**	Hits@10 26.9036 RHD **22.17**	Hits@10 25.4695 RHD **25.13**
MART	Hits@10 27.0836 RHD **21.82**	Hits@10 24.3816 RHD **26.44**	Hits@10 23.3667 RHD **29.67**	Hits@10 24.8509 RHD **28.00**	Hits@10 27.8299 RHD **19.49**	Hits@10 25.5026 RHD **25.04**
SemRoDe-MMD	Hits@10 27.3246 RHD **21.13**	Hits@10 25.8912 RHD **21.89**	Hits@10 23.8995 RHD **28.06**	Hits@10 23.8488 RHD **30.91**	Hits@10 27.1320 RHD **21.51**	Hits@10 25.6192 RHD **24.69**
ATFL	Hits@10 27.3119 RHD **21.16**	Hits@10 24.9778 RHD **24.65**	Hits@10 23.3667 RHD **29.67**	Hits@10 24.0137 RHD **30.43**	Hits@10 26.5228 RHD **23.27**	Hits@10 25.2386 RHD **25.81**

on MS-HGAT. Even architectures built on non-Euclidean geometry (RotDiff) or continuous-time ODEs (GODEN) suffer significant disruption, with RHDs reaching 21.97% and 22.39%, respectively. This confirms that discrete gradient-guided identity substitutions generated on a single source model generalize reliably across structurally heterogeneous forecasting methods.

Ineffectiveness of mainstream adversarial defenses. Standard adversarial defense mechanisms provide negligible protection against PIP. Standard AT provides the best suppression among evaluated defenses, but it only marginally reduces the average RHD from 25.66% to 22.75% and leaves all the target models severely disrupted. Regularization-based methods such as TRADES (average RHD 25.13%), MART (average RHD 25.04%), SemRoDe-MMD (average RHD 24.69%), and ATFL (average RHD 25.81%) perform almost identically to the undefended baseline. For instance, on DisenIDP, SemRoDe-MMD yields an RHD of 30.91%, which matches the undefended setting exactly, while ATFL increases the RHD on DyHGCN to 24.65% (up from 23.42% undefended). These results highlight a critical limitation of existing defenses: while traditional methods optimize for continuous feature perturbations or topological edge flips, they fail to defend against discrete and behaviorally constrained participant identity substitutions in temporal propagation trajectories.

Architectural vulnerability analysis. The degree of model vulnerability correlates with how architectures process propagation identity representations. Models that construct higher-order group interactions from participant sequences (e.g., DisenIDP and MS-HGAT) are especially vulnerable (RHD $> 29\%$), since replacing key propagation identities alters the underlying hyperedge topologies and disentangled influence channels. Geometric and continuous-time models (RotDiff and GODEN) show slightly lower sensitivity but still lose over 20% of their predictive accuracy under PIP. Overall, the consistent failure of existing defenses across all architectures underscores the urgent need for defense mechanisms tailored specifically to propagation identity security.

6 Conclusion

In this work, we propose Propagation Identity Perturbation (PIP), a new adversarial threat model that formalizes attacks targeting participant identities in temporal propagation trajectories rather than static graph topology or node features. To optimize discrete identity substitutions while preserving realistic propagation behavior, PIP restricts candidate search to a behaviorally plausible space governed by temporal activity consistency, propagation role semantics, and neighborhood compatibility, coupled with a gradient-guided discrete search mechanism. We provide theoretical foundations for PIP by establishing guarantees on feasible search space reduction, transferability lower bounds across target architectures, monotonic loss ascent properties, and guaranteed loss disruption bounds.

Extensive experiments across five representative social event forecasting architectures and five mainstream defense mechanisms show that PIP generates highly transferable attacks that cause an average Relative Hits@10 Drop of 25.66% and consistently bypasses existing defense measures. Our analysis indicates that vulnerability in temporal diffusion forecasting is primarily driven by propagation-stage sensitivity rather than conventional static graph structural sensitivity. These findings expose a significant security blind spot in existing temporal graph models and underscore the urgent need for identity-aware defense mechanisms tailored specifically to propagation dynamics.

REFERENCES

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019.

Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1149–1158, New York, NY, USA, 2017. Association for Computing Machinery.

Zhangtao Cheng, Wenxue Ye, Leyuan Liu, Wenxin Tai, and Fan Zhou. Enhancing information diffusion prediction with self-supervised disentangled user and cascade representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3808–3812, New York, NY, USA, 2023. Association for Computing Machinery.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1115–1124. PMLR, 10–15 Jul 2018.

Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 321–338, Santa Clara, CA, August 2019. USENIX Association.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Brian Formento, Wenjie Feng, Chuan-Sheng Foo, Anh Tuan Luu, and See-Kiong Ng. SemRoDe: Macro adversarial training to learn representations that are robust to word-level attacks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8005–8028, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Simon Geisler, Tobias Schmidt, Hakan Şirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7637–7649. Curran Associates, Inc., 2021.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 577–586, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Hongliang Qiao, Shanshan Feng, Xutao Li, Huiwei Lin, Han Hu, Wei Wei, and Yunming Ye. Rotdiff: A hyperbolic rotation representation model for information diffusion prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2065–2074, New York, NY, USA, 2023. Association for Computing Machinery.

Wenbo Shang, Zihan Feng, Yajun Yang, and Xin Huang. Make information diffusion explainable: Llm-based causal framework for diffusion prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 132830–132857. Curran Associates, Inc., 2025.

Ling Sun, Yuan Rao, Xiangbo Zhang, Yuqian Lan, and Shuanghe Yu. Ms-hgat: Memory-enhanced sequential hypergraph attention network for information diffusion prediction. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 4156–4164, 2022.

Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant Honavar. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *Proceedings of The Web Conference 2020*, pp. 673–683, New York, NY, USA, 2020. Association for Computing Machinery.

Ding Wang, Wei Zhou, and Songlin Hu. Information diffusion prediction with graph neural ordinary differential equation network. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9699–9708, New York, NY, USA, 2024. Association for Computing Machinery.

Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. Adversarial training with fast gradient projection method against synonym substitution based text attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13997–14005, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Chunyuan Yuan, Jiacheng Li, Wei Zhou, Yijun Lu, Xiaodan Zhang, and Songlin Hu. Dyhgcn: A dynamic heterogeneous graph convolutional network to learn users' dynamic preferences for information diffusion prediction. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pp. 347–363, Berlin, Heidelberg, 2020. Springer-Verlag.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations*, 2019.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2847–2856, New York, NY, USA, 2018. Association for Computing Machinery.

A Appendix

A.1 Proofs of Theorems

A.1.1 Proof of Theorem 1

Proof. Let $\mathcal{S}$ be the perturbed positions. For fixed $|\mathcal{S}| = b$, each position has at most m feasible replacements. Therefore, the partial constrained search space has size

$$N_b^{\mathcal{S}} \leqslant \binom{k}{b} m^b. \tag{24}$$

On the other hand, the partial unconstrained search space has size

$$N_b = \binom{k}{b} (|\mathcal{U}| - 1)^b. \tag{25}$$

Since both terms are positive,

$$\frac{N_b^{\mathcal{S}}}{N_b} \leqslant \frac{\binom{k}{b} \cdot m^b}{\binom{k}{b} \cdot (|\mathcal{U}| - 1)^b} = \frac{m^b}{(|\mathcal{U}| - 1)^b} = \left(\frac{m}{|\mathcal{U}| - 1}\right)^b, \quad N_b^{\mathcal{S}} \leqslant \left(\frac{m}{|\mathcal{U}| - 1}\right)^b N_b. \tag{26}$$

Summing up both sides of the right inequality w.r.t. b, we have

$$|\Pi_B^{\mathcal{C}}| = \sum_{b=0}^{B} N_b^{\mathcal{S}} \leqslant \sum_{b=0}^{B} \left(\frac{m}{|\mathcal{U}| - 1}\right)^b N_b \leqslant \sum_{b=0}^{B} N_b = |\Pi_B|. \tag{27}$$

This proves the upper bound of the entire constrained search space in (17).

□

A.1.2 Proof of Theorem 2

Proof. First, we expand Wh_ϑ as follows:

$$W\Delta h_\vartheta = Wh_\vartheta(\mathcal{T}') - Wh_\vartheta(\mathcal{T}). \tag{28}$$

Inserting $h_\theta(\mathcal{T})$ twice, we have

$$W\Delta h_\vartheta = \Delta h_\theta + (Wh_\vartheta(\mathcal{T}') - h_\theta(\mathcal{T}')) - (Wh_\vartheta(\mathcal{T}) - h_\theta(\mathcal{T})). \tag{29}$$

Using the triangle inequality twice, we have

$$\begin{aligned} \|W\Delta h_\vartheta\|_2 \geqslant & \|\Delta h_\theta\|_2 - \|Wh_\vartheta(\mathcal{T}') - h_\theta(\mathcal{T}')\|_2 - \|Wh_\vartheta(\mathcal{T}) - h_\theta(\mathcal{T})\|_2 \\ \geqslant & \|\Delta h_\theta\|_2 - 2\epsilon. \end{aligned} \tag{30}$$

Since W is diagonal, we have

$$\|W\Delta h_\vartheta\|_2^2 = \Delta h_\vartheta^\top W^\top W \Delta h_\vartheta = \Delta h_\vartheta^\top \cdot I \cdot \Delta h_\vartheta = \|\Delta h_\vartheta\|_2^2. \tag{31}$$

Inserting (31) into (30), we have

$$\|\Delta h_\vartheta\|_2 \geqslant \|\Delta h_\theta\|_2 - 2\epsilon. \tag{32}$$

□

A.1.3 Proof of Theorem 3

Proof. At each iteration o, Algorithm 1 computes

$$\Delta\mathcal{L}_o(i, v) = \mathcal{L}(\mathcal{T}_{i \leftarrow v}^{(o)}) - \mathcal{L}(\mathcal{T}^{(o)}).$$

It accepts

$$(i^*, v^*) = \underset{(i,v) \in \mathcal{P}_o}{\arg\max}\, \Delta\mathcal{L}_o(i, v) \quad \text{subject to } \Delta\mathcal{L}_o(i^*, v^*) > 0. \tag{33}$$

Therefore,

$$\mathcal{L}(\mathcal{T}^{(o+1)}) = \mathcal{L}(\mathcal{T}^{(o)}) + \Delta\mathcal{L}_o(i^*, v^*) > \mathcal{L}(\mathcal{T}^{(o)}). \tag{34}$$

□

A.1.4 Proof of Theorem 4

Proof. To simplify notations, we denote $h = h_\vartheta(\mathcal{T})$, $h' = h_\vartheta(\mathcal{T}')$, and $\Delta_H = h' - h$. Using the fundamental theorem of calculus,

$$\begin{aligned}&\mathcal{L}(h') - \mathcal{L}(h)\\ =&\mathcal{L}(h + \Delta_H) - \mathcal{L}(h)\\ =&\int_0^1 \nabla_h \mathcal{L}(h + t\Delta_H)^\top \Delta_H \, \mathrm{d}t. \end{aligned} \tag{35}$$

Subtracting both sides of (35) by $\nabla_h \mathcal{L}(h)^\top \Delta_H$ leads to

$$\begin{aligned}\mathcal{L}(h') - \mathcal{L}(h) - \nabla_h \mathcal{L}(h)^\top \Delta_H &= \int_0^1 (\nabla_h \mathcal{L}(h + t\Delta_H) - \nabla_h \mathcal{L}(h))^\top \Delta_H \, \mathrm{d}t,\\ |\mathcal{L}(h') - \mathcal{L}(h) - \nabla_h \mathcal{L}(h)^\top \Delta_H| &= \left| \int_0^1 (\nabla_h \mathcal{L}(h + t\Delta_H) - \nabla_h \mathcal{L}(h))^\top \Delta_H \, \mathrm{d}t \right|. \end{aligned} \tag{36}$$

By Cauchy–Schwarz inequality, the right side of (36) yields:

$$\begin{aligned}&\left| \int_0^1 (\nabla_h \mathcal{L}(h + t\Delta_H) - \nabla_h \mathcal{L}(h))^\top \Delta_H \, \mathrm{d}t \right|\\ \leqslant &\int_0^1 |(\nabla_h \mathcal{L}(h + t\Delta_H) - \nabla_h \mathcal{L}(h))^\top \Delta_H| \, \mathrm{d}t\\ \leqslant &\int_0^1 \|\nabla_h \mathcal{L}(h + t\Delta_H) - \nabla_h \mathcal{L}(h)\|_2 \cdot \|\Delta_H\|_2 \, \mathrm{d}t\\ \leqslant &\int_0^1 t\zeta \|\Delta_H\|_2^2 \, \mathrm{d}t \qquad (37)\\ = &\frac{\zeta}{2} \|\Delta_H\|_2^2, \qquad (38)\end{aligned}$$

where (37) exploits the ζ-smoothness property in Assumption 3. By exploiting the latent continuity in Assumption 3, we have

$$\begin{aligned}\|\Delta_H\|_2 =&\|h' - h\|_2\\ \leqslant&\mathscr{L}_h \sum_{i=1}^{k} \|E(u_i) - E(\tilde{u}_i)\|_2\\ =&\mathscr{L}_h \sum_{i \in \mathcal{S}} \|E(u_i) - E(u_i')\|_2 \qquad (39)\\ =&\mathscr{L}_h \sum_{i \in \mathcal{S}} \|\Delta_{e_i}\|_2, \qquad (40)\end{aligned}$$

where (39) holds due to the definition $\tilde{u}_i = \begin{cases} u_i', & \text{if } i \in \mathcal{S}, \\ u_i, & \text{otherwise.} \end{cases}$ Combining (36), (38), and (40), we have

$$\begin{aligned}|\mathcal{L}(h') - \mathcal{L}(h) - \nabla_h \mathcal{L}(h)^\top \Delta_H| &\leqslant \frac{\zeta \mathscr{L}_h^2}{2} (\sum_{i \in \mathcal{S}} \|\Delta_{e_i}\|_2)^2,\\ \mathcal{L}(h') \geqslant \mathcal{L}(h) + \nabla_h \mathcal{L}(h)^\top \Delta_H &- \frac{\zeta \mathscr{L}_h^2}{2} (\sum_{i \in \mathcal{S}} \|\Delta_{e_i}\|_2)^2. \end{aligned} \tag{41}$$

Substituting $h = h_\vartheta(\mathcal{T})$ and $h' = h_\vartheta(\mathcal{T}')$ in (41), the theorem is proved. □

A.2 Source Model and Five Target Models

A.2.1 Source Model

To generate transferable propagation identity perturbations under a zero-target-query constraint, we employ a pre-trained Twitter checkpoint of MILD (Shang et al., 2025) as the source model, strictly

selected based on validation performance. For all attack generation runs, the input-side influence bias of MILD is set to zero, which ensures that perturbation optimization relies only on generalizable discrete gradient signals rather than architecture-specific module artifacts. The attack optimization procedure operates independently of any internal MILD-specific structural components.

A.2.2 Target Models and Structural Heterogeneity

To rigorously evaluate cross-architecture transferability, we select five state-of-the-art social event forecasting models as victim targets:

RotDiff (Qiao et al., 2023) is a hyperbolic geometric model designed for social event forecasting and information diffusion prediction. Operating under the premise that hierarchical cascade topologies naturally align with non-Euclidean spaces, it employs hyperbolic rotation transformations to embed user interactions. By learning representations over hyperbolic manifolds, RotDiff captures multi-scale hierarchical structures and dynamic participant interaction patterns directly from an ordered user prefix sequence.

DyHGCN (Dynamic Heterogeneous Graph Convolutional Network, Yuan et al. 2020) is a forecasting framework that jointly models static social network structures and temporal diffusion paths. It projects user social ties and time-stamped cascade event sequences into a unified dynamic heterogeneous graph representation. By applying graph convolutions with integrated temporal encodings, DyHGCN captures evolving user preferences and time-varying influence dynamics across ordered propagation prefixes.

MS-HGAT (Memory-enhanced Sequential Hypergraph Attention Network, Sun et al. 2022) exploits hypergraph neural networks to model higher-order group interactions among cascade participants. Recognizing that traditional pairwise edges cannot fully depict multi-user diffusion dynamics, it performs sequential hypergraph attention through node-to-hyperedge and hyperedge-to-node aggregation. It integrates a memory module to retain global cross-cascade user preferences, which enables next-user ranking from prefix user sequences.

DisenIDP (Cheng et al., 2023) applies disentangled representation learning to social event forecasting by decoupling the underlying factors that drives propagation trajectories. It constructs hypergraph structures from both social graph topologies and cascade events to isolate distinct influence channels, such as user topic preferences and contextual trends. This disentangled factor decomposition allows the model to predict participant sequences while processing ordered user input prefixes.

GODEN (Graph Neural Ordinary Differential Equation Network, Wang et al. 2024) is a continuous-time social event forecasting model designed to overcome the temporal discretization artifacts of traditional snapshot-based methods. It models the continuous co-evolution of user dynamic preferences and relationship ties using a coupled system of neural ordinary differential equations (ODEs) defined over graph nodes and edges. Combined with multi-head context attention, GODEN infers future cascade trajectories from ordered prefix sequences under specific maximum-length sequence handling protocols.

A.2.3 Interface Commonality and Representation Invariance

The established benchmark shows high heterogeneity at the representation layer. The target models share no common latent dimensions, embedding coordinate systems, message-passing rules, or temporal encoding formulations. Their structural commonality is strictly confined to the semantic input interface: the ordered sequence of user participation events and the downstream next-user ranking objective.

A.3 Five Defense Baselines

To rigorously evaluate defense capabilities against Propagation Identity Perturbations (PIP), we adapt five representative adversarial defense objective families to the discrete temporal propagation domain. All defenses are trained on paired clean and adversarial propagation sets without altering the underlying target architectures, legal user vocabulary, social graph topology G, or temporal masks.

Let $\mathcal{T} = \{(u_i, t_i)\}_{i=1}^{k}$ be a clean propagation sequence prefix, $\mathcal{T}' \in \Pi_B(\mathcal{T})$ be the corresponding discretely perturbed propagation generated under the PIP threat model, and Y be the true event propagation targets. Let $p_\vartheta(\cdot \mid G, \mathcal{T})$ be the eventwise predictive distribution parameterized by ϑ. The formal objective formulations for the five defense baselines are detailed below. Note that ϑ denotes the parameters of the **target model**, which are **not** those θ of the **source model**.

A.3.1 Adversarial Training (AT)

AT (Madry et al., 2018) constructs a balanced empirical risk minimization objective over clean propagations and discrete PIP adversarial endpoints:

$$\mathcal{L}_{\text{AT}} = \tfrac{1}{2}\mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}), Y\right) + \tfrac{1}{2}\mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}'), Y\right), \tag{42}$$

where $\mathcal{L}_{\text{CE}}$ denotes the standard multi-class cross-entropy loss over the eligible user space. AT forces the model to learn invariant representations across discrete identity replacements without explicitly regularizing intermediate hidden states.

A.3.2 TRADES

Adapted from Tradeoff Adversarial Defense via Surrogate Loss, TRADES (Zhang et al., 2019) decouples natural accuracy optimization from robust boundary regularization by combining clean empirical cross-entropy with a Kullback-Leibler (KL) divergence penalty between clean and perturbed predictive distributions:

$$\mathcal{L}_{\text{TRADES}} = \mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}), Y\right) + \beta \cdot D_{\text{KL}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}) \| p_\vartheta(\cdot \mid G, \mathcal{T}')\right), \tag{43}$$

where $\beta > 0$ controls the trade-off between clean performance and output distribution invariance under discrete participant modifications.

A.3.3 MART

Misclassification-aware Adversarial Risk Theory (MART, Wang et al. 2020) incorporates explicit misclassification awareness into adversarial optimization. By weighting the KL divergence robustness term according to the clean prediction probability assigned to the true label y, MART prioritizes hard and misclassified samples:

$$\mathcal{L}_{\text{MART}} = \mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}'), Y\right) + \beta \cdot D_{\text{KL}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}) \| p_\vartheta(\cdot \mid G, \mathcal{T}')\right) \cdot \left(1 - p_\vartheta(y \mid G, \mathcal{T})\right). \tag{44}$$

This formulation prevents over-regularizing high-confidence clean predictions while penalizing boundary vulnerability on ambiguous propagations.

A.3.4 SemRoDe-MMD

Semantic Robustness Diffusion via Maximum Mean Discrepancy (SemRoDe-MMD, Formento et al. 2024) enforces latent representational alignment between clean and perturbed diffusion sequences. Let $H(\mathcal{T}) = [h_1, \ldots, h_k]^\top \in \mathbb{R}^{k \times \mathfrak{d}}$ be the intermediate temporal event representations extracted before the final projection head. SemRoDe-MMD minimizes:

$$\mathcal{L}_{\text{MMD}} = \mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}), Y\right) + \lambda_{\text{MMD}} \cdot \mathcal{D}^2_{\text{MMD}}\left(H(\mathcal{T}), \mathbf{H}(T')\right), \tag{45}$$

where $\mathcal{D}^2_{\text{MMD}}$ measures the squared Maximum Mean Discrepancy in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ equipped with a universal Gaussian kernel:

$$\mathcal{D}^2_{\text{MMD}}\left(H(\mathcal{T}), H(\mathcal{T}')\right) = \left\| \frac{1}{k}\sum_{i=1}^{k} \phi(h_i) - \frac{1}{k}\sum_{i=1}^{k} \phi(h'_i) \right\|^2_{\mathcal{H}_k}. \tag{46}$$

$\phi(\cdot)$ denotes the canonical feature mapping into the RKHS $\mathcal{H}_k$, and $\|\cdot\|_{\mathcal{H}_k}$ denotes the induced norm via the inner product of $\mathcal{H}_k$. This penalizes trajectory-level semantic shifts caused by discrete identity substitutions.

A.3.5 ATFL

Adversarial Training with Feature and Logit Alignment (ATFL, Wang et al. 2021) combines discrete endpoint adversarial cross-entropy with explicit logit consistency regularization. Denote $z(G, \mathcal{T}) \in \mathbb{R}^{|\mathcal{U}|}$ as the raw predictive logits prior to softmax normalization, the ATFL objective is formulated as:

$$\mathcal{L}_{\text{ATFL}} = \tfrac{1}{2}\mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}), Y\right) + \tfrac{1}{2}\mathcal{L}_{\text{CE}}\left(p_\vartheta(\cdot \mid G, \mathcal{T}'), Y\right) + \gamma \cdot \|z(G, \mathcal{T}) - z(G, \mathcal{T}')\|_2^2, \quad (47)$$

where $\gamma > 0$ regularizes output logit variances to ensure smooth decision boundaries under local identity perturbations.

A.4 Experimental Setup

A.4.1 Dataset

To evaluate defense robustness under realistic operational constraints and eliminate evaluation bias, we establish a standardized defense benchmark by the Twitter corpus distributed with the DisenIDP framework (Cheng et al., 2023). The corpus is partitioned into 2763 training, 345 validation, and 346 test cascades following a strict identity-preserving split protocol. To ensure a rigorous and fair evaluation of defense strategies against PIP, all global context variables are strictly frozen throughout data processing, defense training, and evaluation, including raw-user index mappings, social graph topology G, legal user vocabulary $\mathcal{U}$, and chronological cascade ordering. Defense models are trained on paired clean and perturbed training trajectories generated under valid candidate constraints, which ensures zero feedback or data leakage from validation or test splits.

Evaluation is conducted at event-level granularity to measure fine-grained next-user ranking recovery. The test set contains exactly 7883 non-padding target prediction events for RotDiff, DyHGCN, MS-HGAT, and DisenIDP. As for GODEN, sequence truncation under its official maximum-length handling protocol yields 7880 evaluation events. This strict alignment guarantees that defense performance metrics are compared across identical prediction horizons.

A.4.2 Common Implementation and Evaluation Protocol

Strict attack and defense isolation. To enforce rigorous experimental control and guarantee zero-target-query transferability, attack generation operates strictly on clean Twitter cascades, the shared legal user vocabulary, clean event labels, and source model parameters. The attack algorithm operates with zero access to target model checkpoints, parameters, or defense outputs. Once generated, the perturbed training and test cascades are permanently frozen prior to target defense training. While target defenses may operate under a white-box assumption relative to their own target model during defensive training, the benchmarked attack remains entirely target-independent and defense-unaware.

Standardized defense protocol. Each defense model is initialized from its respective baseline architecture and trained with the complete set of 2763 clean–attacked training pairs. To isolate the defense capability of each defense objective, all core model components are held strictly fixed, including the target architecture, legal user vocabulary, graph/hypergraph construction rules, and evaluation masks. Checkpoint selection is governed strictly by validation cascade performance, incorporating epoch 0 as a no-update fallback, after which clean and perturbed test endpoints are evaluated exactly once. All defense models are trained for 10 epochs at most with an early stopping patience of 3, a learning rate of 2×10^{-5}, and a weight decay of 10^{-5}. Corresponding training batch sizes for RotDiff, DyHGCN, MS-HGAT, DisenIDP, and GODEN are 32, 32, 16, 40, and 8, respectively. We set $\beta = 6$ for TRADES and MART, and $\lambda_{\text{MMD}} = 10$ for SemRoDe-MMD. DyHGCN, MS-HGAT, DisenIDP, and GODEN use AdamW (Loshchilov & Hutter, 2019) for optimization, while RotDiff uses RiemannianAdam (Becigneul & Ganea, 2019).

For every experimental configuration, we evaluate three directly interpretable endpoints:

- **Clean:** Performance of the undefended base model evaluated on clean test cascades.
- **Undefended:** Performance of the undefended base model evaluated on the frozen perturbed test cascades.

- **Defended:** Performance of a specific defense evaluated on the frozen perturbed test cascades.

Evaluation metrics and registered configurations. All primary attack and defense metrics report Event-micro Hits@10 percentages, which pools all non-padding next-user events before computing the hit rate:

$$\text{Hits@10} = \frac{1}{|\mathcal{E}|} \sum_{\mathfrak{e} \in \mathcal{E}} \mathbb{I}\left[y_{\mathfrak{e}} \in \text{Top10}(f_\vartheta(\mathcal{T}'_{1:k_e}))\right], \tag{48}$$

where $\mathcal{E}$ denotes the event set. Pooling events prevents a short cascade and a long cascade from receiving the same weight only because they are separate rows. All clean, undefended, and defended values in the formal tables use this same aggregation rule. Graph and hypergraph structures are constructed from training split data, and defense checkpoints are selected strictly without observing test-set performance.

The source-side attack generation fixes the perturbation budget to $B = 6$, maximum proposal iterations to $R = 60$, gradient candidate set size to 4 per mutable position, and exact source candidates to at most 128 per iteration prior to target evaluation. Target model outputs are never queried or exploited to tune any of these settings.

Plausible zero-query transferability via shared observational semantics. While target forecasting architectures employ fundamentally distinct representational paradigms (ranging from hyperbolic rotations in RotDiff to continuous-time neural ODEs in GODEN), they operate over a shared discrete observation interface defined by the ordered sequence of raw participant identities $\mathcal{T}_{1:k} = \{(u_i, t_i)\}_{i=1}^{k}$. Perturbing a participant identity $u_i \to u'_i$ alters the underlying historical sequence from which downstream graph structures, dynamic attention kernels, and temporal state transitions are synthesized. As a result, our threat model evaluates zero-query attack transferability through shared observational semantics at the raw input layer rather than relying on geometric alignment in a common latent embedding coordinate system.

A.5 Additional Experiments

A.5.1 Ablation Analysis

To show that the attack effectiveness of PIP stems from principled gradient-guided optimization rather than arbitrary candidate noise, we conduct an ablation study on the representative MS-HGAT architecture. Specifically, we compare the default PIP framework (combining gradient proposal with exact verification under behaviorally plausible candidate constraints) against two variants: the Clean baseline and the Pure Random Legal Replacement. The latter randomly substitutes mutable participant identities with candidate users drawn uniformly from the plausible candidate set $\mathcal{C}_i$ under the identical perturbation budget $B = 6$. Results in Table A1 reveal three insights that validate the attack design of PIP.

Table A1: Ablation Analysis for PIP.

Ablation	MS-HGAT
Clean	Hits@10 33.2234%
Pure Random Legal Replacement	Hits@10 28.7581% RHD 13.44%
Gradient Proposal + Exact Verification (Default)	Hits@10 **23.3667%** RHD **29.67%**

Vulnerability of diffusion dynamics to plausible identity perturbations. Even unguided pure random legal replacements in the plausible candidate set induce a notable drop in forecasting accuracy (Hits@10 dropping from 33.22% to 28.76%, yielding an RHD of 13.44%). This confirms our core hypothesis: social event forecasting models depend intrinsically on the precise identities of temporal diffusion participants, which makes participant identities an inherently fragile attack surface even when substitutions follow temporal, propagation, and neighborhood plausibility.

Superior efficacy of gradient-guided search. While random legal substitution exerts moderate disruption, coupling gradient proposals with exact source-model verification significantly amplifies attack severity. The full PIP attack achieves an RHD of $\mathbf{29.67}\%$ (reducing Hits@10 to 23.37%), which is more than **double** the impact of random replacement. This performance gap indicates that PIP does not rely on injecting random variance into diffusion sequences. Instead, the first-order gradient score effectively navigates the discrete candidate space and pinpoints high-impact participant substitutions that maximize loss disruption.

Optimization efficiency in plausible candidate space. The ablation highlights that PIP achieves an optimal balance between behavioral plausibility and adversarial effectiveness. By restricting optimization to the candidate set $\mathcal{C}_i$, PIP avoids generating ineligible or anomalous trajectory artifacts while exploiting first-order sensitivity vectors to locate the most destructive identity swaps in the reduced search space $\Pi_B^{\mathcal{C}}$. This confirms PIP as a mathematically grounded and highly targeted threat model capable of uncovering systemic vulnerabilities in temporal social event forecasting systems.

A.5.2 Hyperparameter Sensitivity Analysis

We conduct hyperparameter sensitivity analysis for four main hyperparameters of PIP with their default settings shown as follows: modification budget $B = 6$ in (3), proposal steps $R = 60$ in Algorithm 1, temporal tolerance $\rho = 3$ in (4), and hop distance $r = 3$ in (6). We change one hyperparameter in a local neighborhood of its default setting, while keeping other hyperparameters fixed at their default settings. The source model MILD (see Appendix A.2.1) itself is used in this hyperparameter sensitivity analysis. B and R are analyzed on the test set, because they serve to benchmark threat model performance under varying capacity limits on unseen target models. ρ and r are analyzed on the validation set, because they belong to the constraint search and adjustment stage. The Clean Hits@10, Attacked Hits@10 by PIP, Absolute Hits@10 Drop (pp), and RHD are reported in the following tables.

Modification budget B. We vary $B \in \{1, 2, \ldots, 8\}$ and show the results in Table A2. As B increases from 1 to 8, the forecasting performance deteriorates monotonically, with the Attacked Hits@10 dropping continuously from 31.447% to 21.223% and the RHD increasing from 9.06% to 38.63%. Remarkably, even at an extremely tight budget of $B = 1$ where the adversary replaces only a single user identity in the entire temporal propagation prefix, PIP induces an absolute performance drop of 3.133 percentage points (pp) and an RHD of 9.06%. This highlights that temporal diffusion forecasting models show non-trivial vulnerability even under minimal and highly localized participant perturbations.

The marginal attack disruption is most severe in the lower budget interval ($B = 1 \to 4$), where RHD grows rapidly from 9.06% ($B = 1$) to 16.21% ($B = 2$), 20.87% ($B = 3$), and 27.00% ($B = 4$). Beyond $B = 4$, the rate of degradation shows slight diminishing returns as the most critical influence positions are already perturbed, but performance continues to decline steadily. At $B = 6$, PIP achieves an Attacked Hits@10 of 22.796%, which corresponds to an absolute drop of 11.785 pp and an RHD of 34.08%. Pushing B further to 8 yields an Attacked Hits@10 of 21.223% (13.358 pp drop, 38.63% RHD). We select $B = 6$ as the default budget for the main experiment as it strikes a favorable trade-off between stealthiness (altering at most 6 identities) and severe adversarial disruption.

Table A2: Hyperparameter sensitivity analysis of PIP w.r.t. B on the test set.

Setting	Clean Hits@10 (%)	Attacked Hits@10 (%)	Absolute Drop (pp)	RHD (%)
$B = 1$	34.581	31.447	3.133	9.06
$B = 2$	34.581	28.974	5.607	16.21
$B = 3$	34.581	27.363	7.218	20.87
$B = 4$	34.581	25.244	9.337	27.00
$B = 5$	34.581	24.052	10.529	30.45
$B = 6$	34.581	22.796	11.785	34.08
$B = 7$	34.581	22.022	12.559	36.32
$B = 8$	34.581	21.223	13.358	38.63

Proposal steps $R = 60$**.** We vary $R \in \{57, 58, 59, 60, 61, 62, 63\}$ and report the results in Table A3. The attack effectiveness across all tested values of R is invariant. Specifically, across all settings, the Clean Hits@10 remains constant at 34.581%, the Attacked Hits@10 remains at 22.796%, the Absolute Drop is identical at 11.784 pp, and RHD maintains 34.08%. It indicates that setting $R = 60$ as the default iteration threshold is sufficient to ensure convergence and provide a reliable margin that avoids premature termination.

Table A3: Hyperparameter sensitivity analysis of PIP w.r.t. R on the test set.

Setting	Clean Hits@10 (%)	Attacked Hits@10 (%)	Absolute Drop (pp)	RHD (%)
$R = 57$	34.581	22.796	11.785	34.08
$R = 58$	34.581	22.796	11.785	34.08
$R = 59$	34.581	22.796	11.785	34.08
$R = 60$	34.581	22.796	11.785	34.08
$R = 61$	34.581	22.796	11.785	34.08
$R = 62$	34.581	22.796	11.785	34.08
$R = 63$	34.581	22.796	11.785	34.08

Temporal tolerance ρ**.** We evaluate the sensitivity of PIP w.r.t. $\rho \in \{1, 2, 3, 4, 5, 6, 7\}$ days, which governs the temporal window size for constructing plausible candidate sets $\mathcal{C}_i^{(t)}$ in (4). As shown in Table A4, PIP performs stably across varying ρ while consistently causing severe performance deterioration on the forecasting system.

At a restrictive setting of $\rho = 1$ day, PIP achieves its highest attack disruption and reduces Hits@10 from 39.425% to 25.348%, which corresponds to an absolute drop of 14.077 pp and an RHD of 35.71%. When relaxing the temporal constraint to $\rho = 2$ days, the Attacked Hits@10 shifts slightly to 25.485% (13.941 pp drop, 35.36% RHD). From $\rho = 3$ days onward, the attack effectiveness reaches a clear saturation plateau: the Attacked Hits@10, Absolute Drop, and RHD remain identical at 25.709%, 13.717 pp, and 34.79%, respectively, across all $\rho \in \{3, 4, 5, 6, 7\}$ days.

This performance plateau reflects the structural composition of the behaviorally plausible candidate space $\mathcal{C}_i$ in ((8)). At $\rho \geqslant 3$ days, the temporal candidate set $\mathcal{C}_i^{(t)}$ becomes sufficiently broad to cover nearly all active users within the target propagation window. As a result, the candidate set intersection is primarily constrained by propagation consistency $\mathcal{C}_i^{(p)}$ and graph neighborhood bounds $\mathcal{C}_i^{(n)}$, rather than the temporal tolerance window ρ itself. These empirical findings indicate that PIP is robust to the specific choice of ρ, and choosing $\rho = 3$ days as the default configuration effectively enforces temporal activity alignment without compromising adversarial optimization or trajectory plausibility.

Table A4: Hyperparameter sensitivity analysis of PIP w.r.t. ρ on the validation set.

Setting	Clean Hits@10 (%)	Attacked Hits@10 (%)	Absolute Drop (pp)	RHD (%)
$\rho = 1$ days	39.425	25.348	14.077	35.71
$\rho = 2$ days	39.425	25.485	13.941	35.36
$\rho = 3$ days	39.425	25.709	13.717	34.79
$\rho = 4$ days	39.425	25.709	13.717	34.79
$\rho = 5$ days	39.425	25.709	13.717	34.79
$\rho = 6$ days	39.425	25.709	13.717	34.79
$\rho = 7$ days	39.425	25.709	13.717	34.79

Hop distance r**.** We vary $r \in \{1, 2, 3\}$ and show the results in Table A5, which governs the structural reach of the neighborhood consistency constraint $\mathcal{C}_i^{(n)}$. r defines the maximum topological distance on the social graph G from the previous participant u_{i-1} within which candidate user replacement v are sampled. Under a strict 1-hop constraint ($r = 1$), candidate substitutions are tightly localized to direct neighbors or immediate transition successors. Even under this highly

restricted candidate pool, PIP causes substantial performance deterioration and reduces Hits@10 from 39.425% to 29.664%, which corresponds to an absolute drop of 9.761 pp and an RHD of 24.76%. Increasing the search radius to a 2-hop neighborhood ($r = 2$) significantly boosts attack potency, drives the Attacked Hits@10 down to 26.274% (13.151 pp drop), and increases the RHD to 33.36%.

Increasing the neighborhood distance further to $r = 3$ yields a minor marginal gain in attack disruption, brings the Attacked Hits@10 to 25.709% (13.717 pp drop), and reaches an RHD of 34.79%. This performance trajectory shows that a 2- to 3-hop neighborhood already encompasses the vast majority of structurally influential and plausible identity replacement candidates. Hence $r = 3$ is selected as the default parameter setting to achieve significant adversarial disruption while enforcing realistic social neighborhood bounds.

Table A5: Hyperparameter sensitivity analysis of PIP w.r.t. r on the validation set.

Setting	Clean Hits@10 (%)	Attacked Hits@10 (%)	Absolute Drop (pp)	RHD (%)
$r = 1$	39.425	29.664	9.761	24.76
$r = 2$	39.425	26.274	13.151	33.36
$r = 3$	39.425	25.709	13.717	34.79

AI USE STATEMENT

In this work, we used generative AI tools only for grammar correction and minor language edits to improve readability. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.