**Goal:**
1. To explore the weakness of temporal graph neural networks (TGNs) by proposing an adversarial attack for link prediction on continuous-time dynamic graphs (CTDGs).
2. To enhance the robustness of TGNs against perturbations and alleviate performance degradation.

**Proposed Algorithms:**
- **T-SPEAR:** a gray-box and poisoning attack for link prediction on CTDGs.
- **T-SHIELD:** a robust training method for TGNs.

**Contributions:**
- We formulate adversarial attacks on CTDGs under 4 realistic constraints regarding unnoticeability.
- Our attack can significantly degrade TGNs’ performances in link prediction.
- Our defense can accurately classify the adversarial edges and mitigate performance degradation.

---

**Spear and Shield: Adversarial Attacks and Defense Methods for Model-based Link Prediction on Continuous-Time Dynamic Graphs**

**Summary**

- Goal:
  1. To explore the weakness of temporal graph neural networks (TGNs) by proposing an adversarial attack for link prediction on continuous-time dynamic graphs (CTDGs).
  2. To enhance the robustness of TGNs against perturbations and alleviate performance degradation.

- Proposed Algorithms:
  - **T-SPEAR:** a gray-box and poisoning attack for link prediction on CTDGs.
  - **T-SHIELD:** a robust training method for TGNs.

- Contributions:
  - We formulate adversarial attacks on CTDGs under 4 realistic constraints regarding unnoticeability.
  - Our attack can significantly degrade TGNs’ performances in link prediction.
  - Our defense can accurately classify the adversarial edges and mitigate performance degradation.

---

**Proposed Defense: T-SHIELD**

- We proposed a robust training method for TGNs, called **T-SHIELD**.
- **Knowledge:**
  - No information about the attacker.
  - Not even know that the dynamic graph has been corrupted.
- **Goal:**
  - To alleviate performance degradation caused by adversarial attacks.

1. **Edge filtering:**
   - To identify and eliminate potential adversarial edges.
   - **Proposed Algorithm:**
     - **T-SPEAR:** a gray-box and poisoning attack for link prediction on CTDGs.
     - **T-SHIELD:** a robust training method for TGNs.

2. **Temporal smoothness:**
   - To prevent sudden changes of node embeddings.

---

**Challenges**

- **C1** What to connect (for $\forall \nu$)?
- **C2** When to connect (for $\forall \nu$)?
- **C3** How to apply a stealthy attack (for $\forall \nu$)?
- **C4** How to discriminate adversarial edges (for $\forall \nu$)?

---

**Proposed Attack: T-SPEAR**

- We proposed an effective adversarial attack method, called **T-SPEAR**.
- **Constraints:** adhering to the four constraints (C1-C4) for unnoticeability.
- **How:** by injecting adversarial edges into the original edge sequence.
- **Goal:** to degrade the link prediction performance of TGN.

---

**Experiments**

**Attack Performance**

- **Dataset**
  - **TGN:** 80.5 ± 0.66, 63.1 ± 1.1.
  - **DGAT:** 60.0 ± 0.9
  - **UAC:** 42.3 ± 0.4
  - **UCI:** 42.3 ± 0.4
  - **Shanghai:** 42.3 ± 0.4
  - **T-SPEAR:** 42.3 ± 0.4

---

**SPEAR attack**

- **Attack**
  - **Defense**

---

**MRR (Mean Reciprocal Rank) as Perurbation Rate Increases**

- **SPEAR attack**
  - For unnoticeability (see details in the paper)
  - **C1** Perturbation budget
  - **C2** Distribution of time
  - **C3** Endpoints of adversarial edges
  - **C4** Number of perturbations per node

---

**Proposed Defense: T-SHIELD**

- We proposed a robust training method for TGNs, called **T-SHIELD**.
- **Knowledge:**
  - No information about the attacker.
  - Not even know that the dynamic graph has been corrupted.
- **Goal:**
  - To alleviate performance degradation caused by adversarial attacks.

---

**Defense Performance**

- **Under Random attack**
  - **Model:** Wikipedia (Clean: 80.5 ± 0.5)
  - **Attack:** T-SPEAR

---

**Code and Data:** [https://github.com/wooner93/T-spear-shield](https://github.com/wooner93/T-spear-shield)