Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9053

Social Bot Detection via Spatio-Temporal Graph Self-Supervision

Hai Lu. LiangXu Shi*
School of Computer Science and Technology, Qingdao University, Qingdao, China
Email: lhai0704@gmail.com

How to cite this paper: Lu, H., & Shi, L. X. (2026). Social Bot Detection via Spatio-Temporal Graph Self-Supervision. *Journal of Computer Science and Frontier Technologies*, 2(3), 125-132. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9053
Published: 2026-03-18

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

Social bots on online social platforms increasingly exhibit profile camouflage, coordinated group behavior, and evolving interaction patterns over time, which makes static-graph and fully supervised detectors less effective under label scarcity and cross-community shifts. To address this issue, this paper proposes STGSLBot, a spatio-temporal graph self-supervised framework for social bot detection. The method constructs a sequence of dynamic interaction snapshots with fixed temporal granularity and initializes node features by combining account metadata with tweet semantics. In addition, temporal statistical signals are introduced to characterize coordinated structures and dynamic reciprocal behaviors. At the model level, a temporal dual-encoder is designed to learn topology-driven commonality and attribute-driven uniqueness, while a causal masked temporal self-attention module aggregates historical information under a strict history-only constraint. During pretraining, neighborhood reconstruction, self-feature reconstruction, cross-view semantic consistency learning, and community-based pseudo-label self-training are jointly optimized to enhance cluster-level discrimination and temporal consistency. Finally, the pretrained model is fine-tuned with a small labeled set for binary classification. Experiments on two temporal subsets derived from TwiBot-22 show that the proposed framework achieves strong and stable performance, especially in accuracy and label-efficient learning.

Keywords

Social Bot Detection; Dynamic Graphs; Self-Supervised Learning; Graph Neural Networks; Spatio-Temporal Representation Learning

1. Introduction

With the rapid expansion of social media and the widespread use of automated content generation tools, social bots have become more influential in misinformation diffusion, opinion manipulation, and fabricated online engagement. Unlike early rule-based bots with rigid behavior patterns, modern bots usually exhibit three coupled characteristics: they imitate human-like profiles and language

styles, coordinate with other accounts at the group level, and continuously evolve their interaction behaviors over time. As a result, bot detection is no longer a purely static classification problem, but a dynamic and structure-dependent task.

Recent graph-based bot detection studies have greatly benefited from large benchmark construction and multi-modal modeling. The release of graph-oriented datasets such as TwiBot-22 has significantly promoted the development of graph-based detectors [1]. Representative methods exploit heterogeneous relations and multimodal user features [2], adaptively distinguish homophilic and heterophilic connections under relation camouflage [3], infer implicit links beyond explicit graph edges [4], or integrate community-aware multimodal features to improve detection robustness [5]. Although these methods enhance representation quality, most of them still focus mainly on static or weakly dynamic social structures.

Another line of research highlights coordinated behavior and temporal evolution. BotCGP shows that coordinated group perception can reveal collective bot patterns that are difficult to capture from isolated account attributes alone [6]. BotDGT further demonstrates that modeling social networks as dynamic graphs and aggregating historical snapshots can improve robustness against evasion strategies [7]. These studies suggest that effective bot detection in realistic scenarios should jointly consider both temporal dynamics and coordinated group behavior.

Meanwhile, label scarcity remains a major bottleneck for practical deployment. High-quality bot annotations are expensive, incomplete, and easy to become outdated. To reduce label reliance, graph self-supervised learning has emerged as a promising solution. Reconstruction-based methods such as GraphMAE2 improve representation robustness through enhanced masked reconstruction [8]. Multi-view contrastive graph learning such as SeBot emphasizes hidden structure and inter-view consistency [9]. In the specific bot detection setting, BotHP further introduces heterophily-aware representation learning and prototype-guided cluster discovery to improve generalization across communities [10]. However, current self-supervised approaches still have two limitations: first, many of them are designed for static graphs or single-time representations; second, instance-level objectives often fail to fully model the cluster-level regularities of coordinated bot groups.

To address these issues, this paper proposes STGSLBot, a spatio-temporal graph self-supervised framework for social bot detection under dynamic and label-scarce settings. The model constructs dynamic interaction snapshots, combines multimodal node features with temporal statistical signals, and learns spatio-temporal representations through a dual-encoder and causal temporal attention mechanism. In the unlabeled stage, multiple pretext tasks and community-based pseudo-label self-training are jointly optimized to improve cluster-level discrimination and cross-time consistency. After pretraining, the model

is fine-tuned on a small labeled set for final classification.

The main contributions of this paper are as follows. First, it proposes a unified spatio-temporal graph self-supervised framework that combines dual-view encoding and causal temporal aggregation within a pretraining-and-fine-tuning paradigm. Second, it introduces temporal statistical signals to capture coordinated structures and reciprocal dynamic behaviors. Third, it enhances unlabeled training with community-based pseudo-label self-training, which improves the separability of coordinated bot groups across time.

2. Model Description

2.1. Problem Definition and Dynamic Snapshot Construction

Given timestamped social interactions and multimodal user information, the observation window is divided into Tconsecutive time intervals to construct a sequence of dynamic graph snapshots:

$$\mathcal{G} = \{G^1, G^2, \dots, G^T\}, G^t = (V^t, E^t). \tag{1}$$

Here, V^tdenotes the active accounts in the t-th interval, and E^tdenotes the observed interaction edges. An edge is added when at least one interaction occurs within the corresponding time slice.

Each node is initialized by combining metadata and tweet semantics. Let $m_i$denote the metadata vector of account i, and let $P_i$denote its historical posts. A pretrained language encoder is used to obtain the text representation:

$$r_i = \frac{1}{|P_i|}\sum_{p\in P_i} \text{RoBERTa}\,(p), x_i = [m_i \parallel r_i]. \tag{2}$$

This multimodal initialization is consistent with recent graph-based bot detection methods that emphasize the importance of user properties and semantic content [2] [5]. The final task is formulated as binary classification, where each account is predicted as either a bot or a human.

2.2. Temporal Statistical Signal Encoding

Only using raw node attributes is often insufficient for discovering coordinated bot groups, while relying on a single snapshot ignores behavioral evolution. Therefore, STGSLBot introduces temporal statistical signals from three aspects: structural density, structural constraint, and reciprocal timing behavior. Specifically, the model computes the k-core level, Burt's structural constraint, and reciprocal response delay for each node in each snapshot. This design is motivated by the observation that community organization and coordinated interaction patterns are crucial to bot detection [6] [5].

The temporal signal embedding for node $i$at time $t$is defined as

$$u_i^t = \left[e_{\text{core}}(c_i^t) \parallel e_{\text{burt}}(b_i^t) \parallel e_{\text{delay}}(d_i^t) \parallel e_t\right]. \tag{3}$$

where $e_{\text{core}}(\cdot)$, $e_{\text{burt}}(\cdot)$, and $e_{\text{delay}}(\cdot)$are learnable embeddings, and $e_t$is an

absolute time embedding used to preserve the global temporal order.

2.3. Temporal Dual-Encoder

In real social networks, bots may exhibit coordinated structural patterns while simultaneously disguising their individual attributes. A single encoder may either over-smooth the representations through excessive neighborhood aggregation or underuse structural information by relying only on node features. To address this issue, STGSLBot adopts a dual-encoder architecture. This design is conceptually related to the distinction between graph-aware commonality and graph-agnostic uniqueness emphasized in BotHP [10].

For each snapshot, a topology encoder based on graph attention is first used to aggregate neighborhood information:

$$h_i^{t,\text{top}} = \text{GAT}(A^t, X^t)_i. \tag{4}$$

At the same time, a self-feature encoder based on a multilayer perceptron preserves node-specific information. The two views are fused as

$$z_i^t = \phi\left(W_f\left[h_i^{t,\text{top}} \parallel \text{MLP}(x_i)\right] + b_f\right). \tag{5}$$

where $\phi(\cdot)$is a nonlinear activation function. The fused representation z_i^tserves as the input to the temporal module.

2.4. Causal Masked Temporal Self-Attention

Dynamic social bot detection should satisfy a strict causality constraint: when making predictions at a certain time, the model should not use future information. To meet this requirement, STGSLBot adopts causal masked temporal self-attention. This design is also aligned with the motivation of dynamic graph modeling in BotDGT, which shows the importance of learning from historical snapshots rather than only the latest graph state [7].

For node i, the fused representations across all snapshots are combined with temporal signal embeddings to form a temporal sequence:

$$\mathbf{H}_i = [z_i^1 + u_i^1, z_i^2 + u_i^2, \ldots, z_i^T + u_i^T]. \tag{6}$$

Self-attention is then applied with a causal mask:

$$\text{Attn}(\mathbf{H}_i) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}. \tag{7}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$are the query, key, and value matrices, and $\mathbf{M}$is the causal mask that blocks each time step from attending to future positions. In this way, the model can capture long-range temporal dependencies without violating temporal causality.

2.5. Self-Supervised Pretraining

To alleviate label scarcity, STGSLBot is first pretrained on unlabeled data. The pretraining stage jointly optimizes four objectives. The first is neighborhood-guided reconstruction, which enhances the structural branch by reconstructing

neighborhood-level information. The second is self-feature reconstruction, which preserves node individuality and reduces over-smoothing. The third is cross-view semantic consistency, which aligns the topology-aware and attribute-aware temporal representations. The fourth is community-based pseudo-label self-training, which uses automatically discovered communities as weak supervision signals to strengthen cluster-level discrimination.

The reconstruction-based design is consistent with the idea of enhanced self-supervised graph learning in GraphMAE2 [8], while the multi-view alignment and structure-aware consistency objectives are related to the motivation of SeBot [9]. In addition, the community-based objective is supported by the findings of BotCGP and BotCF, both of which emphasize the value of coordinated groups and community structure in social bot detection [6] [5].

The overall pretraining loss is

$$\mathcal{L}_{\text{pre}} = \lambda_1 \mathcal{L}_{\text{nbr}} + \lambda_2 \mathcal{L}_{\text{self}} + \lambda_3 \mathcal{L}_{\text{align}} + \lambda_4 \mathcal{L}_{\text{com}}. \tag{8}$$

where $\mathcal{L}_{\text{nbr}}$, $\mathcal{L}_{\text{self}}$, $\mathcal{L}_{\text{align}}$, and $\mathcal{L}_{\text{com}}$denote the neighborhood reconstruction loss, self-feature reconstruction loss, cross-view alignment loss, and community contrastive loss, respectively.

2.6. Fine-Tuning and Inference

After pretraining, the model is fine-tuned with a small labeled set. To obtain an account-level representation, temporal attention pooling is used over the snapshot-level outputs:

$$\alpha_i^t = \frac{\exp\left(q^\top \tanh\left(W_t o_i^t\right)\right)}{\sum_{\tau=1}^{T} \exp\left(q^\top \tanh\left(W_t o_i^\tau\right)\right)}, g_i = \sum_{t=1}^{T} \alpha_i^t \, o_i^t, p_i = \sigma(w^\top g_i + b). \tag{9}$$

where $o_i^t$is the temporal output at time t, $g_i$is the final account representation, and $p_i$is the probability that account $i$is a bot. A smaller learning rate is used for the pretrained encoder and temporal modules, while a relatively larger learning rate is assigned to the classifier head in order to preserve transferable spatio-temporal knowledge and improve adaptation efficiency.

3. Experiments and Results

3.1. Datasets and Temporal Partition

The experiments are conducted on two temporal subsets derived from the graph-based TwiBot-22 benchmark [1]. Twi22-01 covers the period from September 27, 2020 to June 4, 2021, while Twi22-02 covers the period from June 4, 2021 to February 9, 2022. Following a temporal subsetting strategy, users with social behaviors in the corresponding time range are retained to construct the dynamic graph sequence.

To model dynamic social history, each dataset is divided into 500 snapshots with a 12-hour temporal granularity. The training, validation, and test sets are randomly

split at a ratio of 8:1:1. This setting allows us to evaluate bot detection under realistic temporal evolution and limited supervision. Since TwiBot-22 is specifically designed as a graph-based social bot benchmark, it provides an appropriate foundation for evaluating dynamic graph detectors [1].

3.2. Baselines and Metrics

The comparison focuses on four representative baselines: BotCGP, BotRGCN, BothH, and BotICC. These methods represent coordinated-group modeling [6], heterogeneous relational multimodal graph learning [2], adaptive homophily-heterophily learning [3], and implicit connection computation [4], respectively. In addition, recent community-aware or self-supervised directions such as BotCF, SeBot, and BotHP further motivate the design rationale of the proposed framework, although their numerical results are not included in the current comparison table [5] [9] [10].

The evaluation metrics reported in this paper are Accuracy and F1-score. Accuracy measures overall prediction correctness, while F1-score is particularly important under class imbalance because it reflects the trade-off between precision and recall.

3.3. Main Results

Table 1 reports the main comparison results on the two temporal subsets. STGSLBot achieves the best Accuracy on both datasets, reaching 79.33 ± 0.22 on Twi22-01 and 81.31 ± 0.61 on Twi22-02. In terms of F1-score, STGSLBot obtains the best result on Twi22-01 with 58.15 ± 0.74, and remains highly competitive on Twi22-02 with 57.70 ± 1.76, which is only 0.11 lower than the strongest F1 baseline, BothH.

Table 1. Performance comparison of different methods on Twi22-01 and Twi22-02.

Methods	Twi22-01		Twi22-02	
	Accuracy	F1	Accuracy	F1
BotCGP	77.83±0.96	52.16±3.46	52.16±3.46	75.01±2.08
BotRGCN	78.95±0.26	52.26±1.82	52.26±1.82	79.32±0.18
BothH	77.01±0.21	56.47±1.21	56.47±1.21	77.77±1.93
BotICC	74.33±0.17	47.25±0.83	47.25±0.83	80.07±0.36
STGSLBot	79.33±0.22	58.15±0.74	58.15±0.74	81.31±0.61

A closer comparison shows that on Twi22-01, STGSLBot improves Accuracy by 0.38 over BotRGCN and improves F1-score by 1.68 over the strongest prior F1 baseline. On Twi22-02, STGSLBot improves Accuracy by 1.24 over BotICC, while maintaining nearly the best F1-score. These results indicate that the proposed framework is especially effective in improving overall classification reliability under temporal distribution changes.

The performance gains can be explained from three aspects. First, the dynamic snapshot formulation and causal temporal aggregation better capture evolving

behavioral patterns than static or weakly dynamic models, which is consistent with the findings of BotDGT [7]. Second, the dual-encoder structure jointly models structural coordination and individual disguise, which is important in the presence of both homophilic and heterophilic interactions, as also emphasized in BothH and BotHP [3] [10]. Third, the self-supervised pretraining and community-based pseudo-label strategy enhance cluster-level discrimination under limited labels, which is aligned with the motivations of GraphMAE2, SeBot, BotCGP, and BotCF [8] [9] [6] [5].

3.4. Ablation Study

Table 2 presents the ablation results of STGSLBot. The full model achieves 58.15 on Twi22-01 and 57.70 on Twi22-02, which are the best results among all ablation settings. Removing the temporal statistical signal reduces F1-score to 55.23 and 53.81, indicating that structural density and dynamic reciprocal behavior provide useful complementary information beyond raw multimodal features.

Table 2. Ablation study of STGSLBot on Twi22-01 and Twi22-02 in terms of F1-score..

Ablation Settings	Twi22-01	Twi22-02
Full Model	58.15±0.74	57.70±1.76
w/o Statistical Signal	55.23±0.14	53.81±0.33
w/o Dual Encoder	43.66±0.28	49.06±0.72
w/o Self Supervised	35.21±2.23	39.29±1.13
w/o Fine Tuning	44.09±0.57	46.82±0.47

Removing the dual encoder causes a much larger drop, to 43.66 on Twi22-01 and 49.06 on Twi22-02. This result confirms that topology-aware commonality and attribute-aware uniqueness are both necessary for robust bot detection, which is also compatible with the heterophily-aware design intuition of BothH and BotHP [3] [10].

The most severe degradation appears when self-supervised pretraining is removed. In this case, F1-score drops sharply to 35.21 and 39.29. This observation strongly verifies that unlabeled pretraining is crucial in label-scarce settings, and it supports the value of reconstruction and multi-view consistency learning highlighted in GraphMAE2 and SeBot [8] [9]. In addition, removing fine-tuning also leads to a significant decrease, which suggests that pretraining alone cannot fully replace task-specific supervised adaptation.

Overall, the ablation study shows that the performance of STGSLBot depends on the cooperation of four components: temporal statistical signal injection, dual-view representation learning, self-supervised pretraining, and labeled fine-tuning. The consistent trend across both datasets demonstrates the robustness of the proposed framework.

4. Conclusion

This paper proposes STGSLBot, a spatio-temporal graph self-supervised framework

for social bot detection. By combining dynamic snapshot construction, temporal statistical signal encoding, dual-view graph representation learning, causal temporal aggregation, and community-aware self-supervised pretraining, the proposed model effectively addresses the challenges of dynamic camouflage, coordinated group behavior, and label scarcity. Experiments on Twi22-01 and Twi22-02 show that STGSLBot achieves the best accuracy on both datasets and strong F1 performance, while the ablation study further verifies the contribution of each module. These results demonstrate that spatio-temporal self-supervised learning is a promising direction for robust social bot detection.

References

[1] Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., ... & Luo, M. (2022). Twibot-22: Towards graph-based twitter bot detection. Advances in Neural Information Processing Systems, 35, 35254-35269.

[2] Feng, S., Wan, H., Wang, N., & Luo, M. (2021, November). BotRGCN: Twitter bot detection with relational graph convolutional networks. In Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining (pp. 236-239).

[3] Li, S., Qiao, B., Li, K., Lu, Q., Lin, M., & Zhou, W. (2023, October). Multi-modal social bot detection: Learning homophilic and heterophilic connections adaptively. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 3908-3916).

[4] Guo, S., Wang, J., Wang, Z., Yu, G., & Wu, S. (2025). BotICC: enhancing social bot detection through implicit connection computation: S. Guo et al. Multimedia Systems, 31(3), 210.

[5] Liu, F., Li, Z., Yang, C., Gong, D., Liu, F., Ma, R., & Bors, A. G. (2025). BotCF: Improving the Social Bot Detection Performance By Focusing on the Community Features. IEEE Transactions on Network and Service Management.

[6] Qiao, B., Li, K., Zhou, W., Li, S., Lu, Q., & Hu, S. (2025, April). Identifying Bots on Social Media through Coordinated Group Perception. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[7] He, B., Yang, Y., Wu, Q., Liu, H., Yang, R., Peng, H., ... & Zhou, P. (2024). Botdgt: Dynamicity-aware social bot detection with dynamic graph transformers. arXiv preprint arXiv:2404.15070.

[8] Hou, Z., He, Y., Cen, Y., Liu, X., Dong, Y., Kharlamov, E., & Tang, J. (2023, April). Graphmae2: A decoding-enhanced masked self-supervised graph learner. In Proceedings of the ACM web conference 2023 (pp. 737-746).

[9] Yang, Y., Wu, Q., He, B., Peng, H., Yang, R., Hao, Z., & Liao, Y. (2024, August). Sebot: Structural entropy guided multi-view contrastive learning for social bot detection. In Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining (pp. 3841-3852).

[10] He, B., Jiang, X., Wu, Q., Liu, H., Yang, Y., & Liao, Y. (2025, August). Boosting Bot Detection via Heterophily-Aware Representation Learning and Prototype-Guided Cluster Discovery. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 860-871).