Adversarial Socialbots Modeling Based on Structural Information Principles

Xianghua Zeng[1], **Hao Peng**[1], **Angsheng Li**[1,2]

[1] State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2] Zhongguancun Laboratory, Beijing, China
{zengxianghua, penghao, angsheng}@buaa.edu.cn, liangsheng@gmail.zgclab.edu.cn

Abstract

The importance of effective detection is underscored by the fact that socialbots imitate human behavior to propagate misinformation, leading to an ongoing competition between socialbots and detectors. Despite the rapid advancement of reactive detectors, the exploration of adversarial socialbot modeling remains incomplete, significantly hindering the development of proactive detectors. To address this issue, we propose a mathematical **S**tructural **I**nformation principles-based **A**dversarial **S**ocialbots **M**odeling framework, namely **SIASM**, to enable more accurate and effective modeling of adversarial behaviors. First, a heterogeneous graph is presented to integrate various users and rich activities in the original social network and measure its dynamic uncertainty as structural entropy. By minimizing the high-dimensional structural entropy, a hierarchical community structure of the social network is generated and referred to as the optimal encoding tree. Secondly, a novel method is designed to quantify influence by utilizing the assigned structural entropy, which helps reduce the computational cost of SIASM by filtering out uninfluential users. Besides, a new conditional structural entropy is defined between the socialbot and other users to guide the follower selection for network influence maximization. Extensive and comparative experiments on both homogeneous and heterogeneous social networks demonstrate that, compared with state-of-the-art baselines, the proposed SIASM framework yields substantial performance improvements in terms of network influence (up to 16.32%) and sustainable stealthiness (up to 16.29%) when evaluated against a robust detector with 90% accuracy.

Introduction

Socialbots, automated user accounts partly controlled by software, have become indispensable in manipulating public opinion on social media (Ferrara et al. 2016; Subrahmanian et al. 2016). However, socialbots often receive criticism for disseminating false or unreliable information, causing confusion among online communities regarding crucial issues (Das, Lavoie, and Magdon-Ismail 2016; Deb et al. 2019; Aral and Eckles 2019). Differentiating socialbots from regular user accounts is challenging due to their diverse and dynamic behaviors that resemble real-life users' behaviors, including forming appropriate followers and engaging in interactions (Cresci 2020; Arin and Kutlu 2023).

Various methods have been proposed to detect socialbots and prevent their negative impact, including supervised and unsupervised machine learning approaches (Chavoshi, Hamooni, and Mueen 2016; Varol et al. 2017; Yang et al. 2023). Researchers have developed a framework that relies on minimal account metadata to improve scalability and generalization for adequate detection (Yang et al. 2020). Nevertheless, these methods are passive as they wait for socialbot evasion before developing appropriate detection measures (Cresci et al. 2021). Instead of relying on these reactive approaches, researchers explore proactive detection methods using the multi-agent hierarchical reinforcement learning (HRL) mechanism (Le, Tran-Thanh, and Lee 2022), which simulates and understands the adversarial behaviors of socialbots. Within this HRL mechanism, two adversarial objectives of socialbots are formulated: to survive under robust detectors by determining activity types and to maximize network influence by selecting good followers. However, implementing the HRL mechanism poses practical challenges. On the one hand, it is insufficient for the upper-level agent to dynamically determine activity types solely by learning from scratch without considering the entire social network structure. On the other hand, the lower-level agent maintains the local features of all users to select followers, which leads to computational inefficiency.

In this paper, we propose a novel mathematical **S**tructural **I**nformation principles-based **A**dversarial **S**ocialbots **M**odeling framework, namely **SIASM**, to address the challenges above and further develop proactive detection. Firstly, we transform the diverse user nodes and their multi-relational interconnections in the original social network into a unified structure, specifically, a heterogeneous graph. We calculate the structural entropy of this graph to quantify its dynamic uncertainty. Secondly, we minimize the high-dimensional structural entropy to generate an optimal encoding tree representing a hierarchical community structure of social users. Each node in this tree corresponds to a user community, where users of the same community engage in frequent interactions. To enhance the computational efficiency of SIASM, thirdly, we present a new method for quantifying the network influence of each community. This method utilizes the assigned structural entropy of the

corresponding tree node to filter out trivial communities with low influence. Fourthly, we define a conditional structural entropy measure between the socialbot node and each user node, which guides the selection of appropriate followers to maximize network influence. Extensive synthetic and real-life social network experiments are conducted to evaluate our proposed framework's network influence and sustainable stealthiness. Comparative results and analyses demonstrate its performance advantages over state-of-the-art baselines. Furthermore, all source codes and experimental results are available at an anonymous link[1].

In summary, the contributions of our work can be summarized as follows:

- An innovative structural information principles-based framework, called SIASM, is proposed to tackle the challenges of insufficiency and inefficiency in adversarial socialbot modeling.
- A novel method for quantifying network influence using the assigned structural entropy of each user community is presented to effectively filter out uninfluential users and reduce the computational complexity of SIASM.
- A new conditional structural entropy between the socialbot and each user node is defined to guide the follower selection and maximize the network influence of socialbot.
- Our experiments on social networks demonstrate that SIASM achieves significant improvements of up to $116.64(16.32\%)$ and $13.21(16.29)\%$ in network influence and sustainable stealthiness compared to SOTA baselines.

Preliminaries

Social Network Environment

In this work, we model the social network environment by referring to the problem formulations from the ACORN method (Le, Tran-Thanh, and Lee 2022), which includes network representation, diffusion model, and socialbot.

Network Representation. A social network is modeled as a directed multi-relational graph $G_m = (V, \{\mathcal{E}_a\}_{a\in\mathcal{A}})$, where V is the set of vertices[2] representing social users, $\{\mathcal{E}_a\}_{a\in\mathcal{A}}$ is the set of edges representing various social activities $\mathcal{A}$. When a social activity $a \in \mathcal{A}$ occurs between vertices v_i and v_j, we construct a directed edge $e^a_{i,j} \in \mathcal{E}_a$, which signifies the transmission of news from v_i to v_j, thereby v_i influencing v_j.

Influence Diffusion Model. Similar to previous studies on social networks (Jendoubi et al. 2017; Li et al. 2017), we adopt the Independence Cascade Model (ICM) (Goldenberg, Libai, and Muller 2001) to represent the propagation of real-life news or influence in the graph G_m. In the ICM, a specific set of vertices known as initial followers F are active, while the rest remain inactive. At each discrete timestep, each active vertex $v \in F$ has an equal probability p of activating its inactive neighbors $\mathcal{N}(v)$. Once no additional vertices are left to activate, the propagation process concludes (Kamarthi et al. 2020; Li, Lowalekar, and Varakantham 2021). We use the notation $\sigma(G_m, p)$ to denote the number of vertices a piece of news can reach from F through the ICM model.

[1]https://github.com/SELGroup/SIASM

[2]Vertices are defined in the graph, and nodes are in the tree.

Socialbots. In the graph G_m, a socialbot $b \in B$ refers to a vertex that imitates real-life behaviors with the aim of spreading misinformation or low-credible content. Its adversarial objectives mainly consist of two parts: 1) optimizing network influence by selecting good user nodes as followers $F \subset V$ over time, and 2) evading detection and removal by robust bot detectors.

Markov Decision Process

To simultaneously optimize the above objectives, we model the adversarial behaviors of socialbots as a Markov Decision Process (MDP) (Bellman 1957), denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. This process consists of a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition function $\mathcal{P}$, a reward function $\mathcal{R}$, and a discount factor $\gamma \in [0, 1]$. At each timestep, the agent receives an environmental state $s \in \mathcal{S}$ and chooses an action $a \in \mathcal{A}$ based on its policy $\pi(s, a)$, resulting in a new state $s' \sim \mathcal{P}(s, a)$ and a reward $r \sim \mathcal{R}(s, a)$.

Structrual Information Principles

A partition of the vertices set V in a homogeneous graph $G = (V, E)$ is defined as $P = \{P_0, P_1, \dots\}$, where each P_i is a community that serves as a cluster of vertices. And these communities can be further divided into sub-communities in a hierarchical manner. Unlike traditional information entropy used in communication systems (Shannon 1948), Li and Pan (Li and Pan 2016) first proposed structural entropy to quantify the dynamic certainty embedded in complex networks under such hierarchical partitions. In this work, the hierarchical partitions are represented by a tree structure known as the encoding tree.

Encoding Tree. Similar to the previous study (Zeng, Peng, and Li 2023), the encoding tree T of graph G is formally defined as a rooted tree with the following properties: 1) For the root node λ, the set of vertices corresponding to λ is denoted as $V_\lambda = V$. 2) For each leaf node ν, the set consists of a single vertex $v \in V$, represented as $V_\nu = \{v\}$. 3) For each non-root and non-leaf node α, there exists a subset of vertices V_α corresponding with α, and its parent node is denoted as α^-. 4) For each non-leaf node α, we assume the number of its children as L_α and its i-th child as $\alpha^{\langle i\rangle}$, respectively. 5) For each non-leaf node α, all subsets of vertices $V_{\alpha^{\langle i\rangle}}$ are disjointed and set $V_\alpha = \bigcup_{i=1}^{L_\alpha} V_{\alpha^{\langle i\rangle}}$.

One-dimensional Structural Entropy. Without any hierarchical partitioning structure, the dynamic certainty of graph G is measured as the one-dimensional structural entropy and defined as follows:

$$H^1(G) = -\sum_{v\in V} \frac{d_v}{vol(G)} \cdot \log_2 \frac{d_v}{vol(G)}, \tag{1}$$

where d_v is the degree of vertex v and $vol(G) = \sum_{v\in V} d_v$ is the volume of G.

High-dimensional Structural Entropy. An encoding tree T can significantly reduce the dynamic uncertainty of graph G, and the high-dimensional structural entropy measures the remaining uncertainty embedded in G. For each non-root tree node $\alpha \in T$, its assigned structural entropy is

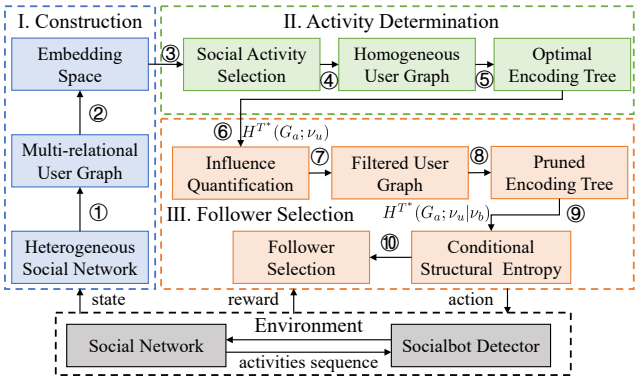

Figure 1: The proposed adversarial socialbot modeling framework: SIASM.

defined as follows:

$$H^T(G;\alpha) = -\frac{g_\alpha}{vol(G)} \log_2 \frac{\mathcal{V}_\alpha}{\mathcal{V}_{\alpha^-}}, \tag{2}$$

where $\mathcal{V}_\alpha$ is the volume of V_α and g_α is the sum of all edge weights connecting each vertex in V_α and each vertex outside V_α. The K-dimensional structural entropy is defined:

$$H^T(G) = \sum_{\alpha \in T, \alpha \neq \lambda} H^T(G;\alpha), \tag{3}$$

$$H^K(G) = \min_T \left\{ H^T(G) \right\}, \tag{4}$$

where T ranges over all encoding trees whose height are at most K, $K > 1$.

The Proposed SIASM Framework

In this work, we adopt the structural information principles to dynamically model the adversarial behaviors of socialbots, as shown in Fig. 1. The SIASM receives the environmental state of the social network, generates a social activity to update the network structure, and obtains reward information for further optimization. Specifically, the architecture of SIASM encompasses three stages: graph construction, activity determination, and follower selection. In the graph construction stage, we transform the original heterogeneous social network into a multi-relational user graph and encode its network structure in an embedding space. In the activity determination stage, we select a social activity to simplify the multi-relational graph into a homogeneous graph and minimize its structural entropy to generate the optimal encoding tree. In the follower selection stage, we quantify the network influence of each user, remove trivial user vertices and tree nodes with low influence from the user graph and encoding tree, and measure conditional structural entropy to guide the follower selection. Fig. 2 showcases the detailed design of the proposed SIASM framework.

User Graph Construction

Originating from the history of social messages, we extract user elements and various social activities, including $tweet$, $retweet$, $mention$, and $reply$, to construct the heterogeneous social network shown in Fig. 2. To preserve heterogeneous information across different elements, we convert the social graph into a multi-relation user graph, denoted as $G_m = (V, \{\mathcal{E}_a\}_{a \in \mathcal{A}})$ in Fig. 2. The vertices V in this multi-relational graph represent social users, including socialbots $B \subset V$ and followers $F \subset V$. These collections are equipped with pre-trained features of word embedding and timestamp encoding, enhancing their semantic representations and temporal information. When users share the same message with different social activities $\mathcal{A}$, edges representing various user relations are established in G_m. As the levels of impurities differ across these relations within the multi-relational graph and collectively impact the embedding results, we adopt the R-GCN network (Schlichtkrull et al. 2018) to encode the structure of G_m. For each user vertex $v \in V$, its pre-trained features and the multi-relational structural information are integrated to encode it into a d-dimensional embedding h_v in Fig. 2, effectively enhancing its representation.

Social Activity Determination

In this stage, SIASM introduces an RL agent that dynamically selects a type of social activity to simplify the multi-relational graph, transforming it into a homogeneous user graph and minimizing its structural entropy. This process generates a hierarchical community structure of social users for subsequent influential follower selection.

At each timestep, the RL agent encodes its activities history as a fixed-dimensional vector and utilizes the learned representation to determine which social activity to perform, aiming for sustainable stealthiness, as depicted in step II. a in Fig. 2. If the $tweet$ is chosen, the influential follower selection stage is skipped within the SIASM framework for that particular timestep. Based on the chosen activity a, we simplify the multi-relational graph G_m into a homogeneous user graph $G_a = (V, \mathcal{E}_a)$ in Fig. 2 by eliminating directed edges representing other types of social activities. To calculate the weight $w_{i,j} \in [-1, 1]$ for each directed edge $e^a_{i,j} = (v_i, v_j)$, we employ Spearman Correlation Analysis on their representations h_{v_i} and h_{v_j} as follows:

$$w_{ij} = 1 - \frac{6 * \left\| k\left(h_{v_i}\right) - k\left(h_{v_j}\right) \right\|_2^2}{d * (d^2 - 1)}, \tag{5}$$

where $k(h_v)$ is the rank of the d-dimensional vertex representation h_v.

To assess the impact of a specific activity a on the social network, we model message propagation as a random walk between users in the homogeneous graph G_a and incorporate structural entropy to quantify the inherent dynamic uncertainty. We generate an optimal encoding tree as a hierarchical partitioning structure of user communities by minimizing its high-dimensional structural entropy. Specifically, we initialize a one-layer encoding tree T for the homogeneous graph G_a as follows: (1) For the entire set of vertices V, we generate a root node λ and set $V_\lambda = V$; (2) For each vertex $v \in V$, we generate a leaf node ν and set $V_\nu = \{v\}$; (3) For each leaf node ν, we assign its father as the root

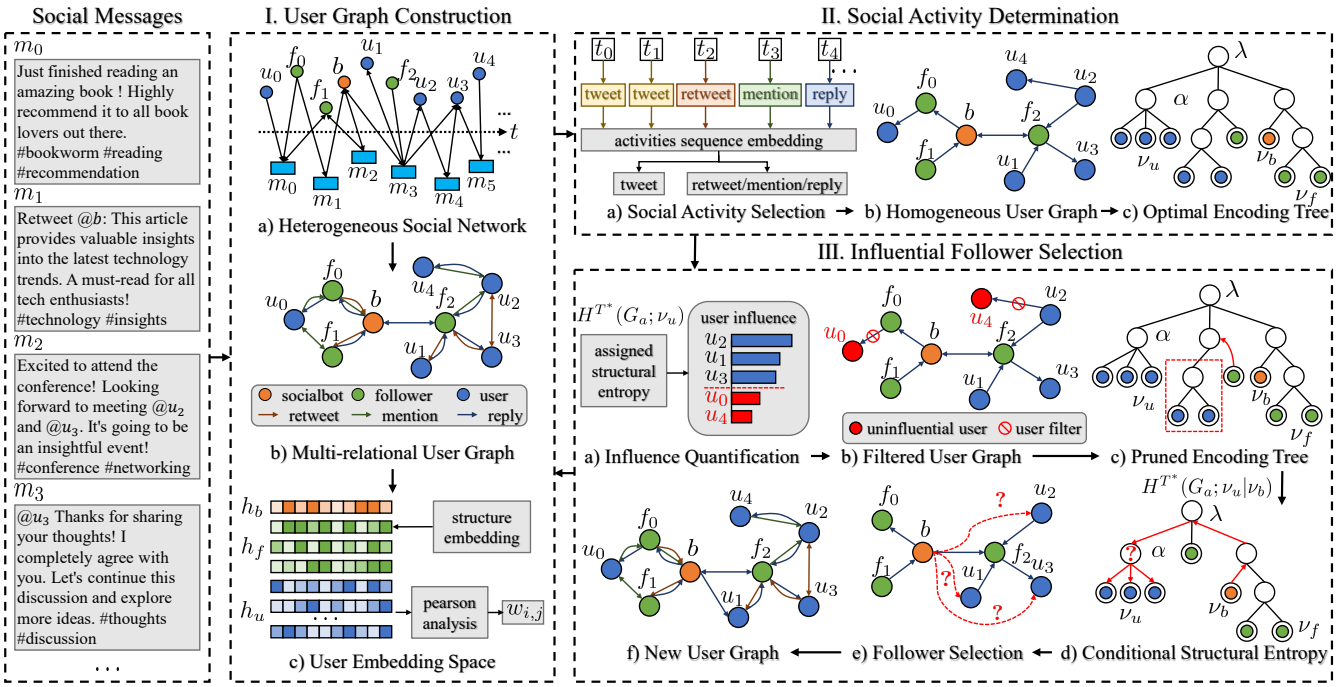

Figure 2: The detailed designs of SIASM framework .

Algorithm 1: The Optimization Algorithm

Input: The one-layer initial encoding tree T, $K \in \mathbb{Z}^+$

Output: The K-layer optimal encoding tree T^*

1 $h_T \leftarrow$ the height of T
2 **while** $h_T < K$ **do**
3 $\quad i^* \leftarrow \arg\max_i\{\overline{R_{se}}(T; U_i)\}$
4 $\quad$ **if** $\overline{R_{se}}(T; U_{i^*}) = 0$ **then**
5 $\quad\quad$ break
6 $\quad$ **for** $\alpha \in U_{i^*}$ **do**
7 $\quad\quad$ *stretch* T_α
8 $\quad\quad$ *compress* T_α
9 $\quad\quad$ $h_T \leftarrow h_T + 1$
10 $\quad$ **for** $i = i^* + 1, ..., h_T$ **do**
11 $\quad\quad$ update U_i
12 $T^* \leftarrow T$
13 **return** T^*

node λ, denoted as $\nu^- = \lambda$. Additionally, we incorporate two operators, *stretch* and *compress*, from the HCSE algorithm (Pan, Zheng, and Fan 2021) to optimize the one-layer encoding tree. In our work, we denote the average reduction of structural entropy resulting from one round of *stretch* and *compress* operations on all tree nodes U_i in layer i as $\overline{R_{se}}(T; U_i)$. We generate the optimal encoding tree T^* with K layers by iteratively and greedily selecting tree nodes to execute the above operations. The optimization process is summarized in algorithm 1. In each iteration, we examine all sets of nodes at the same layer to identify the set U_{i^*} that maximizes the average reduction of structural entropy $\overline{R_{se}}$ (line 3 in algorithm 1) and then execute *stretch* and *compress* operations for each node in U_{i^*} (lines 7 and 8 in algorithm 1). We continue these operations until either the tree height h_T reaches K (line 2 in algorithm 1) or no nodes are satisfying $\overline{R_{se}}(T; U_{i^*}) > 0$ (line 4 in algorithm 1). We terminate the iteration and output T as the optimal encoding tree T^*.

The encoding tree T^* in Fig. 2 illustrates the hierarchical community structure of social users, where each tree node represents a user community, and its height signifies the community's position within the hierarchy. Each leaf node ν corresponds to a singleton consisting of a single user vertex v, with $V_\nu = \{v\}$. Each non-leaf node α corresponds to a new community V_α, which consists of the communities of its children, $V_\alpha = \bigcup_{i=1}^{L_\alpha} V_{\alpha^{(i)}}$. The root node λ corresponds to the entire set of social users, with $V_\lambda = V$.

Influential Follower Selection

After determining the activity type in the previous stage, the SIASM quantifies the network influence of each user community and calculates the conditional structural entropy between the socialbot and each leaf node to guide follower selection for maximizing its social influence.

For each tree node α in the optimal encoding tree T^*, we calculate its assigned structural entropy $H^{T^*}(G_a; \alpha)$ using Eq. 3. Intuitively, a higher value of $H^{T^*}(G_a; \alpha)$ indicates a greater likelihood of a specific type of social activity a occurring between users within community V_α compared to other communities starting at the siblings of α. In step III. a of Fig. 2, we quantify the network influence I_α of user

community V_α by summing up the likelihood of occurrence for each tree node β on the path from the root λ to the node α as follows:

$$I_\alpha = \sum_{V_\alpha \subseteq V_\beta \subset V_\lambda} H^{T^*}(G_a; \beta). \quad (6)$$

To reduce the problem size of follower selection, we prune branches starting at nodes α with low network influence in T^* and filter out all users in the community V_α from the graph G_a, as steps III. b and III. c in Fig. 2. This strategy effectively decreases the number of potential followers to select from, guaranteeing the SIASM framework's efficiency. In this work, we set the ratio of filtered user vertices and the height of pruned subtrees as 5% and 1, respectively.

We trace the path of each potential follower u from ν_u to the root λ and verify if the socialbot b exists in the community V_δ at every tree node δ on this path. When we locate a node δ encompassing both ν and b within its community V_δ, we calculate the conditional structural entropy $H^{T^*}(G_a; \nu_u \mid \nu_b)$ as step III. d in Fig. 2:

$$H^{T^*}(G_a; \nu_u \mid \nu_b) = \sum_{V_{\nu_u} \subseteq V_\alpha \subset V_\delta} H^{T^*}(G_a; \alpha). \quad (7)$$

We leverage conditional structural entropy to measure the uncertainty from the father node δ to the leaf node ν_u, where α is any node on this path. This entropy reflects the probability of a piece of news originating from the socialbot b reaching the user u, and we use it as the initial selection probability for maximizing the socialbot's network influence. To further refine this probability, we employ the actor-critic RL (PPO) algorithm (Schulman et al. 2017). For an efficient optimization process, we utilize the state abstraction mechanism (Zeng et al. 2023) that extracts essential features from all vertices' representation vectors and enables effective follower selection, as step III. e in Fig. 2. Finally, the socialbot adds the user u as a new follower and construct a directed edge of social activity a, thereby yielding an updated user graph in Fig. 2.

Time Complexity of SIASM

In this section, we analyze the time complexity of the SIASM framework, which encompasses user graph construction, social activity determination, and social follower selection stages, to assess its practicality. The overall complexity of SIASM is $O\left(m + n + n \cdot n_a + m \cdot \log^2 n\right)$, where n represents the number of users, m the number of messages, and n_a the number of social activities. Specifically, in the graph construction stage, constructing the social network or multi-relational user graph takes $O(m + n)$ time complexity, and embedding user representations takes $O(m + n \cdot n_a)$ time complexity. In the activity determination stage, simplifying the multi-relational graph incurs a complexity of $O(n_a)$, and optimizing the encoding tree leads to a complexity of $O\left(m \cdot \log^2 n\right)$. During the follower selection stage, the SIASM requires $O(n)$ time complexity to quantify influence, filter user communities, and select appropriate followers.

Experiments and Analysis

In this section, we present empirical and comparative experiments on homogeneous and heterogeneous social networks to validate the superiority of the SIASM framework. And we provide all experimental results, including their corresponding average values and standard deviations. Each experiment is conducted with five random seeds to ensure unbiased evaluations and avoid discrepancies.

Experimental Setup

Datasets. To analyze homogeneous datasets, we collect the top 100 trending articles about the US presidential election and COVID-19 pandemic topics from Twitter and study their propagation networks, which include 1500 social users. For heterogeneous network analysis, we use the latest Higgs Twitter Dataset (De Domenico et al. 2013), which includes directed multi-relational interactions. These networks comprise multiple star-shaped communities with limited connections, which share the same observations as previous research (Sadikov et al. 2011; Kamarthi et al. 2020). Like other works (Le, Tran-Thanh, and Lee 2022), we select 10% of the real-life networks to construct synthetic stochastic networks as the training set and take the remaining 90% of the collected networks as the testing set.

Baselines. This work compares the SIASM with the state-of-the-art adversarial socialbots modeling method ACORN (Le, Tran-Thanh, and Lee 2022). Additionally, we combine several classical heuristic approaches (CELF (Leskovec et al. 2007) and DEGREE (Chen, Wang, and Yang 2009)) with the previously learned agent to create other baselines, known as ACRON-H and SIASM-H.

Evaluations

In this section, we evaluate various methods in both homogeneous and heterogeneous social networks using synthetic graphs to train models and real-life graphs to test their performances. The resulting averages and deviations of episode rewards and lengths are summarized in Table 1. Our analysis shows that SIASM significantly improves episode rewards and lengths in both synthetic and real-life graphs. Specifically, SIASM achieves up to $116.64(16.32\%)$ and $13.21(16.29\%)$ improvements in reward and length, demonstrating its advantages in network influence and sustainable stealthiness. Furthermore, the following subsection separately shows a detailed analysis of experiments conducted on homogeneous or heterogeneous datasets.

Homogeneous Datasets. In each homogeneous social network, we do not consider different types of user activities such as *tweet*, *retweet*, *mention*, and *reply*. Instead, we model them as a homogeneous user graph. During the training process on synthetic graphs, we use a default propagation probability (p) of 0.8 and a maximal episode length (T_{max}) of 120. As depicted in Fig. 3, our SIASM framework converges with fewer environmental steps (312000 and 260000) and achieves better performances regarding network influence (831.36) and sustainable stealthiness (108.45) compared to other methods. These advantages indicate the efficiency of

Table 1: Summary of overall experimental results in homogeneous and heterogeneous datasets: "average value ± standard deviation" and "improvements" (%). **Bold**: the best performance in each graph, underline: the second performance.

Homogeneous	Synthetic Graph		Real-life Graph		Average Performance	
	Episode Reward	Episode Length	Episode Reward	Episode Length	Episode Reward	Episode Length
ACRON-H	-	-	825.53 ± 13.71	40.44 ± 24.48	825.53 ± 13.71	40.44 ± 24.48
SIASM-H	-	-	820.84 ± 11.11	62.78 ± 35.63	820.84 ± 11.11	62.78 ± 35.63
ACRON	714.72 ± 59.49	97.12 ± 8.46	827.96 ± 17.23	67.89 ± 22.62	771.34 ± 38.36	82.51 ± 15.54
SIASM	$\mathbf{831.36 \pm 3.67}$	$\mathbf{108.45 \pm 2.01}$	$\mathbf{831.02 \pm 8.87}$	$\mathbf{81.10 \pm 20.16}$	$\mathbf{831.19 \pm 6.27}$	$\mathbf{94.78 \pm 11.09}$
Abs.(%) Avg. ↑	116.64(16.32%)	11.33(11.67%)	3.06(0.37%)	13.21(16.29%)	5.66(0.69%)	12.27(14.87%)
Heterogeneous	Synthetic Graph		Real-life Graph		Average Performance	
	Episode Reward	Episode Length	Episode Reward	Episode Length	Episode Reward	Episode Length
ACRON-H	-	-	817.56 ± 12.53	74.36 ± 10.42	817.56 ± 12.53	74.36 ± 10.42
SIASM-H	-	-	831.83 ± 7.66	75.39 ± 4.92	831.83 ± 7.66	75.39 ± 4.92
ACRON	800.62 ± 4.08	108.71 ± 1.05	821.83 ± 16.37	71.73 ± 11.36	811.23 ± 10.23	90.22 ± 6.21
SIASM	$\mathbf{828.61 \pm 2.16}$	$\mathbf{110.61 \pm 0.66}$	$\mathbf{840.89 \pm 7.26}$	$\mathbf{80.0 \pm 1.41}$	$\mathbf{834.75 \pm 4.71}$	$\mathbf{95.31 \pm 1.04}$
Abs.(%) Avg. ↑	27.99(3.50%)	1.9(1.75%)	9.06(1.09%)	4.61(6.11%)	2.92(0.35%)	5.09(5.64%)

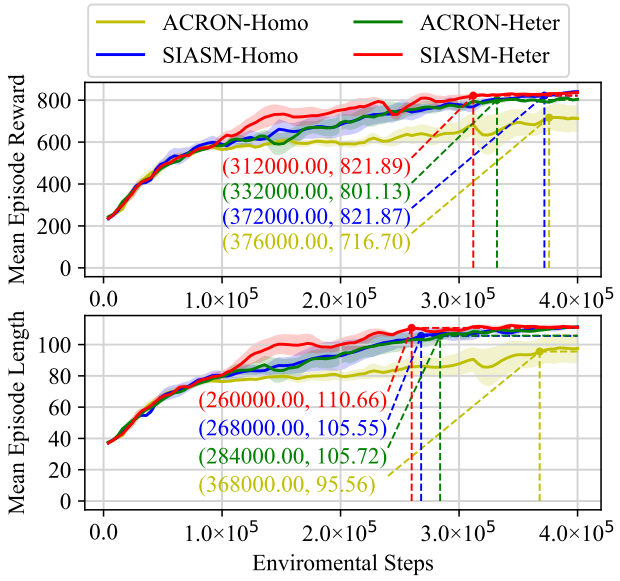

Figure 3: Learning curves of episode reward and episode length in homogeneous and heterogeneous social networks.

SIASM in learning policies for follower selection and detection evasion, leading to desirable outcomes.

During the testing process using real-life graphs, we vary the p-value and measure the network influence ratio, which represents the ratio of users receiving target messages to all users across different follower budgets, as shown in Fig. 4. Overall, SIASM consistently outperforms all baselines, particularly when the number of followers exceeds a threshold of 294, $|F| > 294$. It is evident that the SIASM stably achieves a more significant influence while interacting with fewer social users, regardless of propagation probabilities. This superiority can be attributed to SIASM's ability to select suitable followers based on the global feature of structural entropy, thereby overcoming the limitations associated with baselines that rely on local features for node selection.

Table 2: Total survival timesteps v.s. network influence ratio after reaching $|F| = |V|$. **Bold**: the best performance, underline: the second performance.

Methods	$p = 0.25$		$p = 0.5$	
	% ↑	Steps ↑	% ↑	Steps ↑
ACRON-H	0.20 ± 0.05	$1.2K \pm 1K$	0.37 ± 0.13	$1.2K \pm 1K$
SIASM-H	0.19 ± 0.03	$1.8K \pm 996$	0.68 ± 0.32	$1.1K \pm 1K$
ACRON	0.81 ± 0.34	$2.1K \pm 254$	0.99 ± 0.10	$2.0K \pm 276$
SIASM	$\mathbf{0.99 \pm 0.02}$	$\mathbf{2.4K \pm 147}$	$\mathbf{0.99 \pm 0.03}$	$\mathbf{2.4K \pm 181}$
Methods	$p = 0.25$		average performance	
	% ↑	Steps ↑	% ↑	Steps ↑
ACRON-H	0.53 ± 0.07	$1.2K \pm 1K$	0.37 ± 0.08	$1.2K \pm 1K$
SIASM-H	0.88 ± 0.20	$1.6K \pm 886$	0.58 ± 0.18	$1.5K \pm 961$
ACRON	0.99 ± 0.10	$2.0K \pm 305$	0.93 ± 0.18	$2.0K \pm 278$
SIASM	$\mathbf{0.99 \pm 0.03}$	$\mathbf{2.4K \pm 267}$	$\mathbf{0.99 \pm 0.03}$	$\mathbf{2.4K \pm 198}$

Furthermore, we conduct additional evaluations to assess the sustainable stealthiness of all compared methods. We measure their survival steps in real-life networks and summarize the comparative results in Table 2. The findings reveal that our SISAM demonstrates robust longevity, persisting for $2.4K$ timesteps with a network influence ratio of 0.99 across various real-life scenarios, significantly outperforming other methods.

Heterogeneous Datasets. Under heterogeneous settings, we maintain multi-relational social activities to obtain more comprehensive user embeddings and model adversarial socialbot behaviors based on the principles of structural information, as depicted in Fig. 2. The SIASM and other baselines are trained using the default training parameters, p and T_{max}, which are consistent with homogeneous training. The training curves of these models are presented in Fig. 3, wherein the SIASM exhibits the fastest convergence and achieves the best performance with an episode reward of 828.61 and an episode length of 110.61. By effectively leveraging heterogeneous information from the original social networks, the SIASM and its variant, SIASM-H, demonstrate the highest and second-highest performances in terms of network influence (840.89 and 831.83, respectively) as well as sustainable stealthiness (80.0 and 75.39, respectively) when tested on real-life graphs.

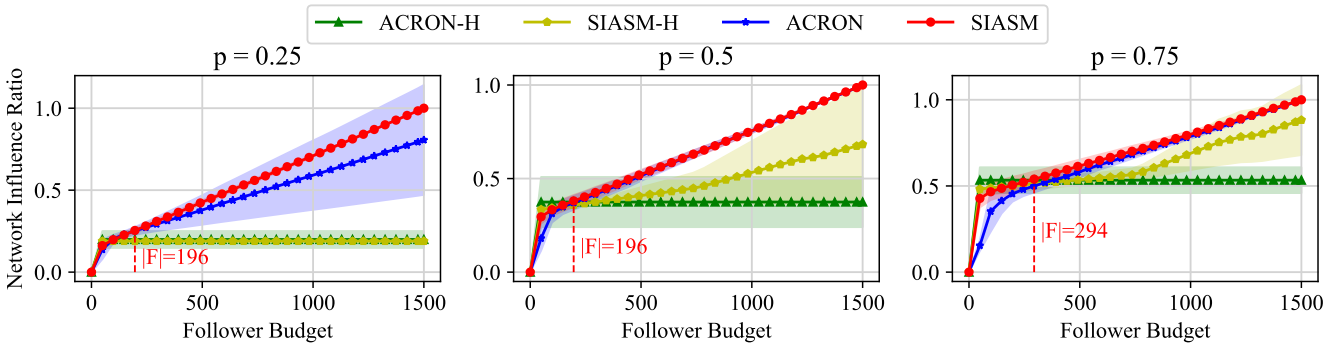

Figure 4: The testing process on homogeneous social networks

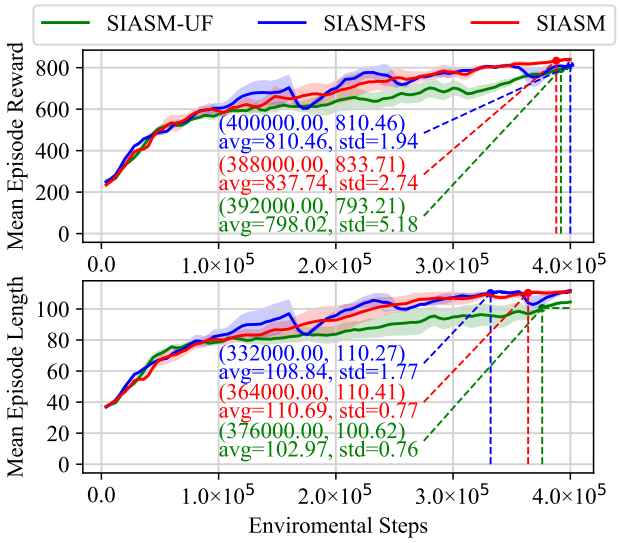

Figure 5: Learning curves of episode reward and length for ablation studies.

Ablation Studies. In this paper, we conduct ablation studies on homogeneous social networks to examine the effects of user filtration and follower selection stages in SIASM. The corresponding variants are referred to as SIASM-UF and SIASM-FS, respectively. The results presented in Fig. 5 demonstrate that removing either the user filtration or follower selection stage decreases the overall quality and efficiency of policy learning. These findings indicate that the filtration and selection stages based on the principles of structural information play a crucial role in enhancing the performance of adversarial modeling.

Related Work

Adversarial Socialbot Modeling (ASM)

Unlike traditional detection methods (Beskow and Carley 2019; Yang et al. 2020) that rely on static snapshots of features, ASM computationally models adversarial socialbot behaviors over time. An evolution optimization algorithm is employed to generate various permutations from a predefined sequence of activities and select the best one to improve the accuracy of detectors (Cresci et al. 2019). However, these permutations only provide static snapshots of behaviors and do not account for the socialbot's evolution. To capture the temporal dynamics of socialbot, a general RL framework (Le, Tran-Thanh, and Lee 2022) formulates adversarial behaviors as a Markov Decision Process (MDP).

Structural Information Principles

The concept of structural information was first introduced in 2016 by Li and Pan (Li and Pan 2016). They proposed a metric that included definitions of structural entropy and partitioning tree, which can measure the dynamic complexity of networks and detect their natural hierarchical structure. The one-dimensional structural entropy minimization principle was then used to identify subtypes of cancer cells by constructing cell sample networks (Li, Yin, and Pan 2016). Later, Li et al. (Li et al. 2018) decoded topologically associating domains of Hi-C data by minimizing high-dimensional structural entropy. In 2023, Zou et al. (Zou et al. 2023) developed SE-GSL, which leveraged structural information principles to enhances the GNN models' robustness towards noisy and heterophily structures. Recently, our team defined state and action abstractions on the encoding trees to achieve efficient and effective general decision-making frameworks (Zeng, Peng, and Li 2023; Zeng et al. 2023).

Conclusion

This paper proposes a structural information principles-based framework SIASM for adversarial socialbots modeling to advance proactive detection. To maximize network influence under robust detectors, an influence quantification method and a conditional structural entropy are designed to guide follower selection. Evaluations of challenging tasks within homogeneous and heterogeneous social networks demonstrate that SIASM significantly improves network influence and sustainable stealthiness compared to state-of-the-art baselines. In the future, we plan to extend our existing work by incorporating adversarial modeling of multiple socialbots and expanding proactive detection.

Acknowledgments

The corresponding authors are Hao Peng and Angsheng Li. This work is supported by National Key R&D Program of China through grant 2022YFB3104700, NSFC through grants 61932002 and 62322202, Beijing Natural Science Foundation through grant 4222030, and the Fundamental Research Funds for the Central Universities.

References

Aral, S.; and Eckles, D. 2019. Protecting elections from social media manipulation. *Science*, 365(6456): 858–861.

Arin, E.; and Kutlu, M. 2023. Deep learning based social bot detection on twitter. *IEEE Transactions on Information Forensics and Security*, 18: 1763–1772.

Bellman, R. 1957. A Markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Beskow, D. M.; and Carley, K. M. 2019. Its all in a name: detecting and labeling bots by their name. *Computational and mathematical organization theory*, 25: 24–35.

Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016. Debot: Twitter bot detection via warped correlation. In *ICDM*, volume 18, 28–65.

Chen, W.; Wang, Y.; and Yang, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 199–208.

Cresci, S. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10): 72–83.

Cresci, S.; Petrocchi, M.; Spognardi, A.; and Tognazzi, S. 2019. Better safe than sorry: an adversarial approach to improve social bot detection. In *Proceedings of the 10th ACM Conference on Web Science*, 47–56.

Cresci, S.; Petrocchi, M.; Spognardi, A.; and Tognazzi, S. 2021. The coming age of adversarial social bot detection. *First Monday*.

Das, S.; Lavoie, A.; and Magdon-Ismail, M. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)*, 10(4): 1–25.

De Domenico, M.; Lima, A.; Mougel, P.; and Musolesi, M. 2013. The anatomy of a scientific rumor. *Scientific reports*, 3(1): 1–9.

Deb, A.; Luceri, L.; Badaway, A.; and Ferrara, E. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. In *Companion proceedings of the 2019 world wide web conference*, 237–247.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12: 211–223.

Jendoubi, S.; Martin, A.; Liétard, L.; Hadji, H. B.; and Yaghlane, B. B. 2017. Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems*, 121: 58–70.

Kamarthi, H.; Vijayan, P.; Wilder, B.; Ravindran, B.; and Tambe, M. 2020. Influence Maximization in Unknown Social Networks: Learning Policies for Effective Graph Sampling. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 575–583.

Le, T.; Tran-Thanh, L.; and Lee, D. 2022. Socialbots on Fire: Modeling Adversarial Behaviors of Socialbots via Multi-Agent Hierarchical Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*, 545–554.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429.

Li, A.; and Pan, Y. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6): 3290–3339.

Li, A.; Yin, X.; and Pan, Y. 2016. Three-dimensional gene map of cancer cell types: Structural entropy minimisation principle for defining tumour subtypes. *Scientific Reports*, 6: 1–26.

Li, A.; Yin, X.; Xu, B.; Wang, D.; Han, J.; Wei, Y.; Deng, Y.; Xiong, Y.; and Zhang, Z. 2018. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Communications*, 9: 1–12.

Li, D.; Lowalekar, M.; and Varakantham, P. 2021. CLAIM: Curriculum learning policy for influence maximization in unknown social networks. In *Uncertainty in Artificial Intelligence*, 1455–1465. PMLR.

Li, M.; Wang, X.; Gao, K.; and Zhang, S. 2017. A survey on information diffusion in online social networks: Models and methods. *Information*, 8(4): 118.

Pan, Y.; Zheng, F.; and Fan, B. 2021. An Information-theoretic Perspective of Hierarchical Clustering. *arXiv preprint arXiv:2108.06036*.

Sadikov, E.; Medina, M.; Leskovec, J.; and Garcia-Molina, H. 2011. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 55–64.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, 593–607. Springer.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Subrahmanian, V. S.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. The DARPA Twitter bot challenge. *Computer*, 49(6): 38–46.

Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, 280–289.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, 1096–1103.

Yang, Y.; Yang, R.; Peng, H.; Li, Y.; Li, T.; Liao, Y.; and Zhou, P. 2023. FedACK: Federated Adversarial Contrastive Knowledge Distillation for Cross-Lingual and Cross-Model Social Bot Detection. In *Proceedings of the ACM Web Conference 2023*, 1314–1323.

Zeng, X.; Peng, H.; and Li, A. 2023. Effective and Stable Role-Based Multi-Agent Collaboration by Structural Information Principles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10): 11772–11780.

Zeng, X.; Peng, H.; Li, A.; Liu, C.; He, L.; and Yu, P. S. 2023. Hierarchical State Abstraction Based on Structural Information Principles. In *IJCAI*.

Zou, D.; Peng, H.; Huang, X.; Yang, R.; Li, J.; Wu, J.; Liu, C.; and Yu, P. S. 2023. SE-GSL: A General and Effective Graph Structure Learning Framework through Structural Entropy Optimization. In *Proceedings of the ACM Web Conference 2023*, 499–510.