Simulating online behaviours and threat patterns for training against influence operations (WiP)

Ulysse Oliveri[1,3,*], Alexandre Dey[2] and Guillaume Gadek[3]

[1] *IRISA, Inria, CNRS, Univ Rennes, Rennes, France*
[2] *Airbus Defence and Space Cyber Programmes, Rennes, France*
[3] *Airbus Defence and Space, Elancourt, France*

Abstract

Social media platforms have enabled large-scale influence campaigns, designed by threat actors to manipulate public opinion. These campaigns use coordinated accounts to spread fake information or amplify information (e.g., disinformation, astroturfing), swaying opinions and paralysing decision-making. To mitigate these impacts, nongovernmental and governmental entities train in simulated informational environments emulating social network platforms and their exchanges. During the trainings, the animation team must implement specific informational Tactics Techniques and Procedures (TTPs) to achieve customized educational objectives.

The simulation of TTPs requires credible social networks, which must notably contain diverse user types (bots, trolls, casual users, influencers, etc.) and recreate social interactions to generate both normal behaviours and malicious behaviours.

This paper introduces a framework designed to generate personalized social networks graphs for training sessions, tailored specifically to the needs of the trainers. This framework allows the modelling of referenced influence operations in order to reproduce specific attacks such as astroturfing or corrupted influencers to increase the training credibility, the pedagogical impact, and capitalising on existing knowledge. We illustrate the coherence of these simulations through two case studies, which aim at reproducing astroturfing attacks and corrupt influencers tactics. We show that our simulation of these tactics coherently reproduces the documented attacks, and we assess the results through topology metrics and information diffusion metrics.

Keywords

Training, Information campaigns, Adversarial Simulation, Social Networks

1. Introduction

The democratisation of social media platforms during the 21st century have enabled large-scale influence campaigns, designed by threat actors to manipulate the public opinion. These campaigns use coordinated accounts to spread fake information or to greatly amplify information operations (e.g disinformation campaigns, astroturfing), in order to sow chaos, and to paralyse decision-making.

Attackers have been continuously improving their methods to take advantage of online social networks, greatly boosting their effectiveness and impact. In response, defenders have organised their Tactics, Techniques, and Procedures (TTPs) through frameworks, such as the DISARM framework [1].

Moreover, in an effort to mitigate and detect these campaigns, entities such as journalists (e.g. fact-checking service), brand monitoring services, company security teams, and government agencies are following training sessions to stay ahead and effectively combat influence operations.

In this context, an animation team (trainers) creates a scenario that depicts the educational goals of the training and decides which targeted informational strategies (i.e. TTPs) need to be implemented in a controlled setting. These TTPs are to be detected and mitigated by the player team (trainees).

To effectively simulate these techniques, it is essential to replicate a credible social network ecosystem. The ecosystem should encompass a variety of user types, such as bots, casual users, influencers, and trolls. Furthermore, the simulation should also accurately model the mechanisms of information diffusion, in order to shape how the diverse topics discussed in a platform interact with each others, and to enable the

C&ESAR'25: Computer & Electronics Security Application Rendezvous, Nov. 19-20, 2025, Rennes, France
*Corresponding author.
ulysse.oliveri@irisa.fr (U. Oliveri)

[1] https://disarmframework.herokuapp.com/

possibility to have viral topics. This modelling is crucial to provide an understanding of the mechanisms exploited by the red team (attackers) for their attacks.

In this paper, we describe a framework that aims to simulate documented attacks such as astroturfing, botnets attacks, or butterfly attacks [2], increasing training sessions likelihood and credibility. The framework generates credible social network graphs along a customizable information diffusion model. The social graph is created taking into account user parameters such as network composition (distribution between user types), density of the network (how much the users are connected), and interactions probability matrix (how often a type of user interacts with another). We summarize our contributions as follows:

- We use the DISARM Matrix to identify useful Tactics, Techniques and Procedures (TTPs) for training sessions, enhancing its educational impacts and providing feedback on the topological effects and changes on the information diffusion caused by the TTPs.
- To simulate these attacks, our framework generates a customizable social graph with end-user inputs such as user account types, network density, and community linkages. It then simulates information diffusion on the generated graph, modelling interactions between regular users and malicious accounts on micro-blogging platforms like X and Mastodon.
- We validated our approach by simulating two TTPs and measured their impact on the social network graph from both topological and information diffusion perspectives. Furthermore, we also verified the impacts of the attacks on the trending hashtags panel of a self-hosted Mastodon instance, to verify the coherence of our simulation with a real social network recommendation system.

2. Related Work

2.1. Social Graph Generation

Reproducing credible social network dynamics is essential to provide users with an adequate training platform. However, creating these graphs is a non-trivial task. Social graphs are composed of various types of users exhibiting diverse online behaviours. The users differ in how they connect with each other [1], in their online activities such as post frequency and reaction frequency [2], and their temporal interaction patterns (i.e., the hours or days at which they interact in the social network) [3]. The reasons for interaction also vary significantly [4]. Bots, for instance, may repeatedly send identical messages to automatically promote a product or flood a network [3]. Trolls might target others on specific subjects to undermine their opponents [5]. Meanwhile, influencers might aim to market products to their followers.

In observed social graphs, the distribution of user connections (i.e., the number of followers) follows a power-law function [6]. This characteristic implies that a small number of users are highly connected, while the majority have relatively few connections.

This phenomenon, often referred to as the "long tail" distribution, means that a few nodes (users) dominate the network's connectivity, while most nodes have limited connectivity. This power-law distribution is a fundamental property of most real-world networks, including online social networks: it significantly shapes their structure and dynamics.

2.2. Information Diffusion in Social Graphs

Information diffusion in social graphs describes the process by which information spreads through a network of interconnected users, driven by user behaviours and network topology.

On this task, literature is usually organised around two approaches, Predictive and Explanatory models[7].

Predictive models aim at predicting the future state of the network after a spark of information appears in the graph. For example, predicting the weight of a piece of information, after a specific user has shared it, is useful for monitoring the network [7]. Foundational work include the Independant Cascade Model [8], which tries to predict the propagation probability of a piece of information in the graph, or the topic-aware model [9] that relates the probability of the propagation to the topic being spread.

[2] https://disarmframework.herokuapp.com/technique/134/view

The explanatory branch includes works trying to mimic nature in itself, with epidemic models [10] such as the SIR (Susceptible-Infected-Recovered) model [11], with the analogy of illness being information, and forest fire models [12], which mimics the process of information diffusion akin to how fire spreads through a forest. In this analogy, the fire represents information, and the trees symbolise the users. Information spreads sequentially from a user to their neighbours (or followers), transitioning from states *tree* to *burning* (active in the graph) to *burnt* (refuses to share information).

When a user reaches a defined probability threshold for information spreading, they transition to the state of *burning*, i.e, the user becomes active. Conversely, if the threshold is not met, the user adopts the state of *burnt*, indicating that they do not further propagate the information.

Modelling information diffusion grants the possibility to emulate informational events such as new topics emerging in the simulated world, and to reproduce influence operation techniques during the training sessions.

2.3. Influence operations techniques

Attackers have developed various operating procedures to enhance their efficiency and impact by leveraging online social networks. Defenders model these operating procedures through Tactics, Techniques, and Procedures (TTPs). Diverse frameworks organise these TTPs, which are diverse in which attackers they want to model [13], what level of information they provide [14, 15] or what thematic they focus (e.g, cognitive techniques) [16]. In this context, we focus on the DISARM framework, which has the right level of granularity over the procedures that need to be reproduced in a training session.

Using the framework, we group the relevant tactics as follows:

- **Massive Content Creation**: This involves coordinated inauthentic behaviours[3], astroturfing [17], and disinformation campaigns [18].
- **Exploitation of Recommendation Systems**: Attackers increase their reach by promoting divisive content and employing emotional triggers to manipulate user engagement [19].
- **False Authority**: Fake accounts claiming to be subject experts (e.g., epidemiologists during COVID-19) to manipulate opinion or cause confusion[4] or corrupted influencers.

Among these classes, tactics such as the massive use of inauthentic accounts often exhibit specific temporal activity [17], with accounts engaging at suspect times compared to genuine behaviour. Moreover, these accounts are used in a structured manner, showing coordination and planning by an attacker.

Disinformation operations significantly alter the structure and dynamics of the social graph, including distorting patterns of influence, artificially inflating the visibility of certain narratives, and disrupting the organic flow of information. These changes may allow defenders to focus their means on important attacks, and metrics analysing in real-time the network are a necessity.

As an example, some metrics measures the topological density of certain communities, detecting abnormal "co-follows, co-retweet, and co-favourite" networks in short time spans[5]. These abnormalities are detected using metrics such as the betweenness centrality of the accounts, their degree centrality, and the exploitation of communities detection and qualification [20], which are or can be used to detect signs of coordination between these accounts.

Threat actors also coordinate in time, interacting in the social network within short time frames [17]. Measuring this gives defenders additional clues on coordination and/or automation. From the attacker point of view, this coordination is needed to manipulate the platform's recommendation system, which increases the visibility of "influential" contents [21] (i.e., here, with high numbers of favourites and retweets in a short time span). This aims to amplify the threat actor's narrative exposure to their target audiences.

Measuring this exposure is quite important, as spending high amounts of time and resources to mitigate failed or limited attacks may not be worth it, due to the limited defender resources. To assess the full impact of the attacks within social networks, a combination of these metrics is required.

[3] https://transparency.meta.com/policies/community-standards/inauthentic-behavior/
[4] https://www.disinfo.eu/publications/disinformation-self-proclaimed-experts-spreading-covid-19-disinformation-under-the-guise-of-expertise/
[5] https://disinfo-prompt.eu/posts/4utNbmaC1keX9Z60IYkxG9

3. Contribution

This paper relies on a controllable social network generation framework, taking as input user parameters such as the user account type repartition, intra-community density and how connected are each community. The framework also features an information diffusion model, integrating key variables representing the user experience on social networks. Finally, we illustrate the framework usefulness by highlighting the possibility (1) to reuse prior documented attacks and their TTPs to capitalise on existing knowledge and (2) to enhance the immersiveness and credibility of training sessions.

3.1. Influence operations tactics to simulate

Our research focuses on tactics that not only alter the topological structure of social graphs but also manipulate the information space. By leveraging the DISARM framework, two critical tactics have been identified that are pivotal to reproduce with our framework:

- **Establish Social Assets**: This tactic involves the deployment of coordinated inauthentic behaviours, including the creation of bots, trolls, or malicious communities. These entities are strategically used to infiltrate existing networks through methods such as massive engagement or butterfly attacks.
- **Establish Legitimacy**: This involves the fabrication of fake experts or opinion leaders to exert influence over communities. Additionally, it encompasses techniques such as corrupting influencers within genuine communities, and astroturfing to create a false consensus, thereby swaying public opinion.

These attacks fundamentally alter the social graph's topology by creating new, often deceptive, relations or connections between communities. Moreover, the behaviour of the accounts involved in these coordinated attacks deviates significantly from that of the normal accounts [17], highlighting a gap in the conventional information dissemination literature.

Our framework aims to bridge this gap by integrating insights from documented attacker behaviours (how they connect, at what time they interact) [17]. By recreating specific attacks used in informational campaigns, we provide the community with a robust tool for capitalising on structured knowledge about attackers' TTPs.

This approach not only enhances the credibility of the training sessions but also equips defenders with a deeper understanding of the potential threats. In order to reproduce these techniques, tactics, and procedures in a dynamic social graph, the framework relies on a graph generation module that produces special user types configurations (trolls, bots, influencers...) controlled by the trainer. These distributions aim at producing tailored user communities to produce specific setups used in the training.

3.2. Social Graph Generation

3.2.1. Community Generation

In platforms such as X (formerly Twitter), the connection between users (nodes of the graph) is defined by the "follow" relationship. This relationship may or may not be reciprocal, significantly impacting the dissemination of information within the network.

In fact, users tend to see more content shared by the accounts they follow, shaping their information exposure and interaction patterns. Furthermore, interactions on X include posting (where a node emits a text post, initiating a thread), replying (contributing to the thread), retweeting (sharing the post or the reply with followers), and favouring (indicating support or approval to a post or a reply).

In observed social graphs, the distribution of user connections (i.e., the number of followers) follows a power-law function [6], which means that a small portion of the users (also called the influencers) concentrate the main proportion of the followers.

In this context, our framework first defines power-law functions parameters for each type of user identified across literature [1, 3, 4, 2]. These types, each with different level of impact on the graph include:

- **Influencers** (Luminary, celebrities, expert, opinion leaders, trendsetters, bloggers, potential influencers).
- **Casual users or Consumers**.

- **Trolls** (Left-troll, Right-troll, Fearmongers, News Feed, Hashtag Gamer).
- **Bots** (Spam Bots, Content Bots, Engagement Bots).

Once these power-laws are defined, the framework proceeds as follows:

1. **Follow probabilities**: A probability matrix of dimension $n_{user\_type} * n_{user\_type}$ is created, according to the source users types and target users types.For example, casual users have a low probability to follow bots accounts, or casual users have a high chance to follow an influencer.
2. **Follower assignation**: Each node iteratively follows others with a given edge probability. Then, v edges are sampled, where v is drawn from a power-law distribution.
3. **Alignment with end-user density parameter**: Random edges are pruned to match the user-specified community density, defined as the ratio of existing edges to all possible edges in a complete graph.
4. **Topic assignation**: A topic distribution - a set of pre-defined topics each associated with a scale from 0 to 1 indicating a user's interest in that topic - is assigned to all nodes. While the topics remain consistent across all nodes, the level of interest for each topic is set-up for individual communities, with small variations from node to node belonging to the same community.

The system applies this algorithm for generating multiple sets of nodes sharing the same topics, called here *communities* (as seen in Figure 2a). Malicious communities and accounts such as the ones used in astroturfing, or butterfly attacks are created using the same algorithm.

An illustration of this community generation module can be seen in Figure 2a where the framework produces 3 communities (0, 1, 2) of 100 nodes each with a diverse user distribution (see Figure 1).

In real social networks, users have the ability to engage with new topics that may not be of interest for their community. In addition, to reproduce the tactics used by attackers (e.g. astroturfing), it is crucial to develop methods to connect separate communities within the network.

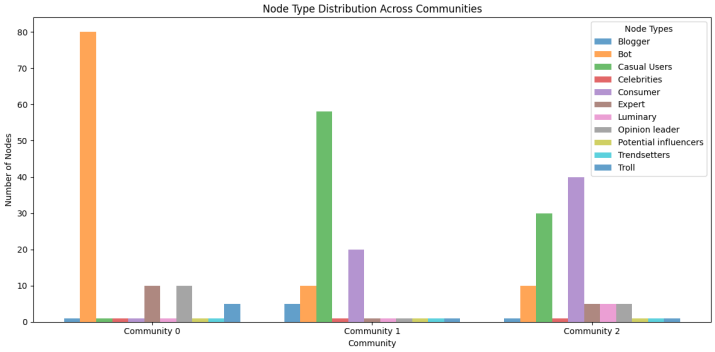


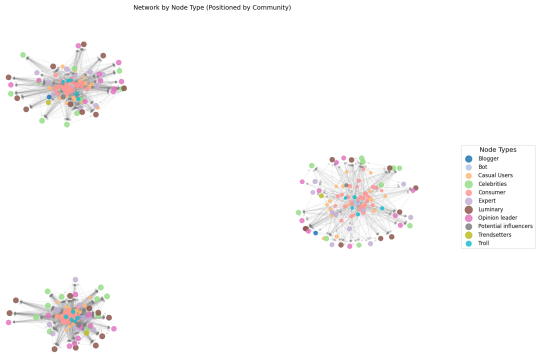
Figure 1: Example of a user type repartition across three generated communities.

3.2.2. Bridges between communities

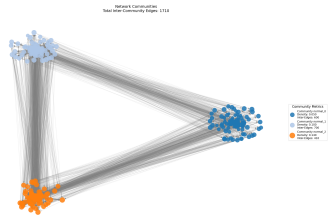
Understanding what are the topics amongst diverse communities, how they interact and who are the main contributors is essential to detect influence operations.

In our system, users can define the **inter-community density** between different communities. This parameter dictates the proportion of connections or edges that exist between two distinct communities, with respect to the maximum connections possible in a complete graph.

To attribute these edges, the system adheres to the same rules that were used during the generation of individual communities. These rules involve assigning relationships based on **user types**, as detailed in the community generation process (see Section 3.2.1). Following this approach, users from different communities become interconnected. This linkage facilitates the **diffusion of information** across the entire parametrised graph, enabling a dynamic and interactive network structure.



(a) Example of generated communities of users, coloured by user type.



(b) Same communities as in Subfigure 2a. End-user parametrises the communities 0, 1, and 2 to connect with specific inter-density parameters. It triggers the creation of bridges between communities, enabling accounts interacting over other communities' topics of interest.

Figure 2: Social network graph creation, initialising with three distinct communities connected later by the user.

In our example, the end-user configures an inter-community matrix that establishes connections between communities, using one to five percent of the total possible edges from one community to another (see Figure 2b). This allows diverse communities to influence each others through their following accounts, mimicating the real social networks.

To effectively model a dynamic social network, one needs to reproduce a coherent information diffusion model. This model captures the temporal evolution of interactions and the spread of information among individuals, reflecting real-world behaviours and network dynamics accurately. By incorporating factors such as the frequency of interactions, the influence of key individuals, and the varying strengths of relationships, the model can simulate how information propagates through the network over time. Additionally, it should account for external influences and changing network structures to provide a comprehensive understanding of the underlying mechanisms driving information diffusion in social networks.

3.3. Information Diffusion

In this subsection, we instantiate an information diffusion model, taking into accounts all the past presented variables (user account types, follower network, topic distribution), and new ones such as the circadian pattern (the rhythm of activity during a day, taking into account sleep hours and work hours).

3.3.1. Algorithm variables

Our information diffusion model draws inspiration from the forest-fire adaptation [12] principle, integrating circadian behavioural patterns, topic dynamics, and user control into the simulation. Unlike the original model, where an account becomes "burnt" if it does not share information, our proposal introduces a probability of being burnt based on the number of times a node refuses to share information on a topic. In addition of the forest-fire model, this paper complements the interaction probability between an emitter and receivers (followers) by incorporating several key variables, separated in two categories (i.e, probability of a post (1), and probability of an interaction with a given post (2)):

- **Behaviour Significance (1) (2)**: This describes how likely a particular type of user is to engage or interact within the graph. It highlights the interaction tendencies of different user types.
- **Circadian Rhythm (1) (2)**: This indicates the current time within the simulation, reflecting how time affects user interactions and behaviours. It is used to simulate if the user is connected or not.
- **Interest of the Topic (1) (2)**: This measures the value or relevance of the target user topic within the source user distribution, indicating how interested the source user is in that topic. For root users, the topic interest is sampled from the user distribution.

- **Target Behaviour Significance (2)**: This refers to the type of user that a given user is interacting with. For root users, this value is not present, as they are the originators of interactions.
- **Virality Parameter over the Topic (1) (2)**: This is a user-specific parameter that controls how much a topic influences the social graph. It allows end-users to adjust the impact of certain topics on interactions, to recreate attackers TTPs.
- **Behaviour Similarity with Other Repliers (2)**: If all other repliers are of the same type as the receiver, this similarity boosts the probability of interaction.
- **Topic Similarity with Other Repliers (2)**: If the receiver notices that other repliers have similar interests in a topic, it increases the probability of interaction.

Moreover, this paper incorporates various interaction types commonly seen in real social networks, such as posts, replies, retweets, and favourites. These interactions are assigned probabilities that are specifically tailored to each user type.

3.3.2. Algorithm description

Initialisation The algorithm begins with the end-user specifying a simulation period (e.g., 3 days from August, the 6th), that will be iterated hour by hour over the period during the simulation. When the end-user wants to model a specific TTP, it is possible to start it at a specific time, allowing users to measure the before/after of the simulation.

For our simulation, the framework processes the chosen topics in parallel, essential to model the cross-topic exposure impact. For each topic, our algorithm starts by selecting k influential nodes, based on:
- Topology: degree centrality and betweenness centrality of the node;
- Interest of the topic: the system weights the influence by the node interest over the topic.

These accounts become *active* (burning in forest-fire model), and are then allowed to proceed with content creation.

Post Creation For each step of the simulation, the active nodes over the given topic are selected. Then, for each node, the module sequentially calculates a probability of
1. Being logged-in (based on the circadian rythm), allowing the user to post and see new contents.
2. If the user is logged in, the user has a probability to create a post (based on the variables in subsection 3.3.1).

If there is a post created, a max breadth and max depth of the future conversation to be generated is sampled from the user type α (the same as the one used to sample the number of follows). To clarify, for influential nodes (such as opinion experts, luminaries...), the future conversations will be deeper (chains of conversations) and broader (the number of replies for each reply will be greater).

Engagement Creation For this part, we categorise as engagement each of the interactions belonging to the set {favourite, retweet, reply}.

After a user creates a post, the module has an approach to greedily sample from the entire network to favour interaction diversity. However, for computational reasons, the system removes 98% of the sampled nodes which do not follow the initial poster. This allows users which do not directly follow another user to interact with their content, similar to what happens in real social networks. In this initial step, the *burnt* nodes are also removed.

For each of the remaining nodes, the framework calculates the probability of interaction with the poster. This results in an interaction type (favourite, retweet, reply).

After replying, the algorithm samples a random probability to continue the conversation recursively on this message, provided depth and breadth constraints are still satisfied, which enables back-and-forth exchanges.

Finally, each new interaction in the graph (post, reply, favourite, retweet) updates the user's topic distribution, reinforcing their interest on the given topic, and lowering their interest on others. This variation allows the players to track the effect of continuous exposure to topics.

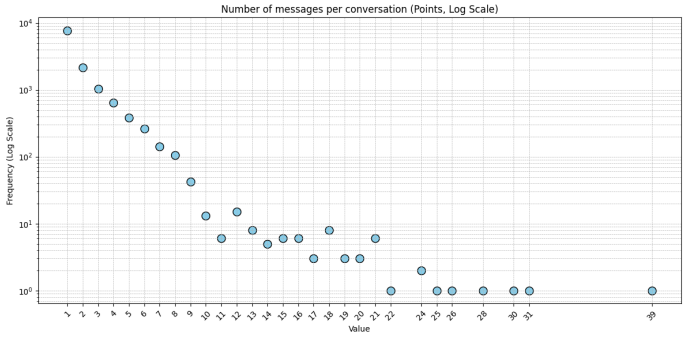


Figure 3: Number of messages per conversation on a three days simulation, with y-axis in log-scale.

As an illustration, Figure 3 shows that, in a three days simulation, the number of messages per conversation follows a power-law distribution akin to what is observed on real social networks such as X [22]. However, the power-law for our simulation is a little bit shifted to the right. In fact, in X, the recommendation system "discards" from the user perspective the messages with zero interactions, and highlights posts with engagement. In our case, the goal is to assess this simulation by deploying it on a self-hosted Mastodon, which does not have a main timeline sorted with a recommendation system. Thus, the goal here is to generate mainly messages with interactions so that our unfiltered Mastodon timeline (a.k.a., live feed) looks similar to the X timeline.

Network updates Social networks are dynamic, with new creations of links between users (new follows), and links being removed (unfollow). To mimic this mechanism, the following heuristics are introduced at each step of the simulation for both actions.

- For follow actions, if user a is exposed to at least $k_{k\in\mathbb{N}}$ posts or replies from user b, and user a has reacted $n_{n\in\mathbb{N}}$ times to them, the user a has shown a non-negligible interest for this user. Hence, at the end of a step, a follow link is created from user a to user b.
- For unfollow actions, if user a has not reacted to the last $p_{p\in\mathbb{N}}$ posts he saw from user b, user a unfollows the user b.

In this section, we presented a framework which allows to recreates a tailored social network experience, creating communities, links between communities, and modelling the diffusion of information.

To demonstrate the credibility of our simulation framework, this paper presents the simulation of two distinct Tactics, Techniques, and Procedures (TTPs).

4. Case Study

In this section, the described system is assessed with the simulation of two examples of frequent Tactics, Techniques, and Procedures (TTPs) documented in the literature.

These TTPs can easily be implemented by an end-user, mainly consists in following six steps:

Steps to generate communities to simulate a Specific TTP using the proposed system

1. **Step 1:** Chose the specific TTP to implement
   `Ex: ttp = "astroturfing"`
2. **Step 2:** Define malicious community distribution and number of accounts to produce.
   `Ex: type_distribution={"Bot":0.2, "Troll":0.1 ...}`
   `total_nodes=500`
3. **Step 3:** Define malicious circadian pattern (probability to be logged in the platform for each hour of the day)
   `Ex: {"0h":0.4, "1h":0.3,"2h":0.5,"3h":0.3 ...}`
4. **Step 4:** Define topic distribution and decide what adversarial narrative to push to the network.
   `Ex: topic_distribution = {"Elections":0.4, "Security":0.2 ...}`
   `narrative = "France should increase military budget ..."`
5. **Step 5:** Define intra and inter-community densities.
   `Ex: Intra-density = 0.3`
   `Inter-density = {"community_0":{"community_1":0.1 ...}`
6. **Step 6:** Decide the time where the attackers start to be active
   `Ex: ttp_start_time = datetime(year=2025, month=8,day=7, hour=12)`

4.1. T0099.001: "Astroturfing"

4.1.1. Sources

This type of attack is widely documented, and thoroughly analysed in [23, 17, 18]. To recreate this attack, we draw inspiration from the analysis of these authors. For example, these papers help us to recreate the interaction patterns of the accounts belonging to an astroturfing campaign [17]. In addition, the proportion of malicious accounts with respect to genuine accounts is also given by [23]. Moreover, these malicious accounts use special semantics, such as the intensive use of keywords in their messages to manipulate recommendation system, as highlighted in Viginum's[6] report.

4.1.2. Technical Instantiation

As presented in the SocialForge data generation system [24], the training setup is initialized with a scenario containing two narratives and two factions, France, and Louraly the attacker.

Reflecting these factions, two account communities are created per faction, with the former having "normal" communities, and the latter having communities belonging to an astroturfing campaign.

The total of generated accounts is 650, where ~20% - 131 accounts - (as observed in [23]) of these are considered malicious. These communities differ on the compositions and the density (astroturfing campaigns have highly-interconnected accounts). The initial density is set to 0.10 for normal communities, and 0.30 for astroturfing communities. Having real-world data on these numbers is quite hard, as having the full landscape of an astroturfing campaign is intractable due to their sheer sizes.

For the composition of the communities, the astroturfing communities contains a greater proportion of bots and trolls, as illustrated in Table 1.

In the training scenario, factions should push narratives, defined as strategic ideas to push to a target audience. For France, the narrative is "France should increase military budget to reduce the threats on its vital interests". For the adversarial communities, it is the contrary, as the narrative is: "French people don't want the French government to increase defence spendings. They should focus on internal problems and not push towards war".

[6] https://www.sgdsn.gouv.fr/files/files/Publications/20250204_NP_SGDSN_VIGINUM_Rapport_public_Elections_roumanie_risques_france_VFF.pdf

Table 1
Distribution of node types and influence scores in communities with and without astroturfing. Astroturfing communities are dominated by **bots (24.2%) and trolls (14.4%)**, while normal communities consist mainly of **casual users (26.2%) and consumers (29.8%)**. Influence scores (average of degree and betweenness centrality) highlight the disproportionate influence of bots and trolls in astroturfing groups, driven by their dense connectivity, whereas **luminaries and celebrities** lead influence in normal communities due to their broad reach.

	Blog.	Bot	Cas.	Celeb.	Cons.	Exp.	Lum.	Opin. Ldr.	Pot. Infl.	Trend.	Troll
Astroturfing (132 nodes)											
Nodes (%)	4 (3.0)	**32 (24.2)**	24 (18.2)	4 (3.0)	27 (20.5)	6 (4.5)	4 (3.0)	4 (3.0)	4 (3.0)	4 (3.0)	**19 (14.4)**
Influence	0.15	**0.07**	0.04	0.23	0.10	0.18	0.28	0.21	0.15	0.14	**0.07**
Normal (531 nodes)											
Nodes (%)	26 (4.9)	26 (4.9)	**139 (26.2)**	26 (4.9)	**158 (29.8)**	26 (4.9)	26 (4.9)	26 (4.9)	26 (4.9)	26 (4.9)	26 (4.9)
Influence	0.15	0.04	0.03	**0.45**	0.08	0.23	**0.43**	0.25	0.13	0.15	0.04

These factions aim at pushing the narratives at all costs, including linking the provided topics to these narratives.

In addition to the topology of the communities and the semantics used in the messages, both communities differ in the time patterns of their activities on social platforms (i.e., different circadian cycles). For example, normal communities follow the natural sleep and work rhythm, with reduced activity during the night and work hours. However, as shown in [17], astroturfing campaigns are especially active during these hours.

To have a comparison ground, and to assess the attack's impact on the social network simulation, the attackers start their campaign only from the middle of the simulation, after one day and twelve hours.

A measure of the coherence of the provided simulation is composed by a set of metrics composed of:

1. Typical social networks metrics such as the distribution of the engagement (number of retweets, favourites, and replies) over time, and topological influence scores.
2. Modularity score before and after the attack, from the Louvain Algorithm. This is used to measure how the communities have merged, showing that the astroturfing are meshing with the normal communities.
3. Narrative exposure to the other communities, measuring how the astroturfing narrative is heard by the other communities.

Furthermore, to validate the simulation system, we retrieve the trending hashtags of a self-hosted Mastodon [7] instance before and after the campaign. This aims at capturing the idea whether the campaign is wide-spread, and whether the recommendation system reflects the manipulation.

In fact, as reported in Viginum's report about Romania's elections, accounts part of an astroturfing campaigns manipulate recommendation systems by posting huge amounts of content that share specificities (e.g., keywords). As such, the astroturfing accounts are provided with a list of ten keywords that should appear later in the "trending" section of Mastodon.

4.1.3. Results

Social network metrics The simulation highlights the differences in circadian activity patterns between normal communities and astroturfing communities, as shown in Figure 4. The key findings are:

- **Normal communities** activity follows a typical circadian rhythm, peaking during standard waking hours.
- **Astroturfing communities**, however, exhibit heightened activity during *off-peak hours* (e.g., late at night or early in the morning), outside of the usual active periods of normal communities.

These results align with the observations of [17], who noted similar discrepancies in activity timing between these groups.

[7] https://joinmastodon.org/fr

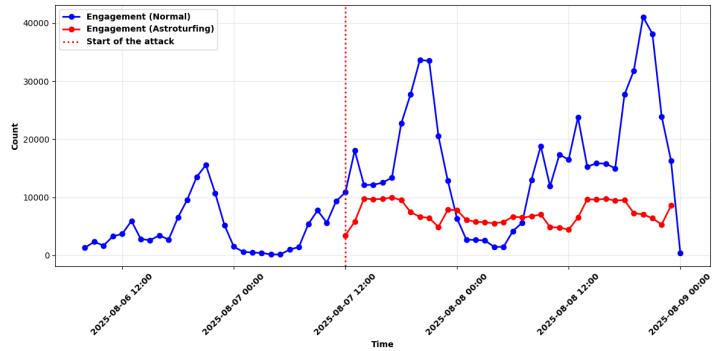


Figure 4: Temporal visualisations of engagement (post, replies, favourites, and retweets) during the simulation, separated by type of communities. Astroturfing communities have a steady level of activation during all the second part of the simulation, with heightened activity during normal communities' off-hours.

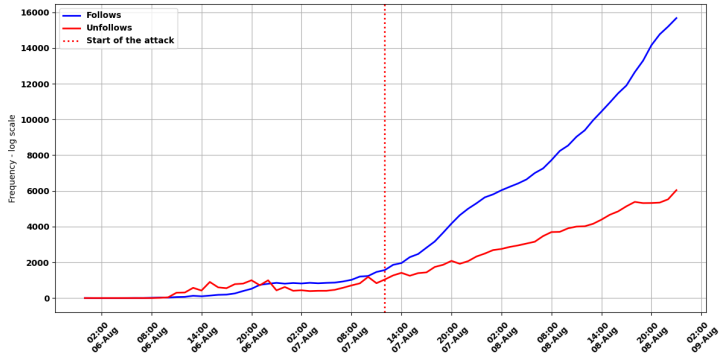


Figure 5: Frequency of follow / unfollow actions across the simulation. Firstly, during the first day of the simulation, there is an increase of follow linked to the start of the simulation. Then, there is a plateau during the second half of the first day. After the beginning of the attack (12:00 the second day), there is a drastic increase of follow actions, adding dynamism to the network and showing that the attackers accounts autonomously start to follow target accounts.

Relation dynamism during the simulation During the simulation, the "follow" relationship is updated based on past interactions with a user. Figure 5 shows that the account follow / unfollow is relatively stable, until the astroturfing attack starts at 12:00 on the second day. After that, there is a drastic increase in the number of follows, with the astroturfing massively following the normal account communities.

Automatic metrics such as **modularity** (given by the Louvain Algorithm) also show a drastic **decrease from 0.60 to 0.20**. This decrease means that communities are less segmented from a topological point of view and there is greater proximity between them.

Narrative exposure One of the metrics an attacker wants to track is how much their adversarial narrative has reached to the target, and in case of an astroturfing campaign, that their target is trapped within an echo chamber over the attackers' narrative, giving the target a false sense of consensus.

This exposure enhances the effectiveness of the attack, aiming at causing kinetic impacts (tangible real-world damage, altering beliefs, eroding public trust, etc.), which grows likelier as the number of reached user increases. Figure 6 illustrates that during the off-hours, normal communities are flooded with messages from the astroturfing communities. As shown in the figure, a large percentage of the interactions are done by or towards astroturfing accounts. During normal hours, we see that these interactions account for 40% to 60%, which is consistent with a short time frame attack goal.

Moreover, as shown in Table 7a, the simulation influenced the trending hashtags, making their message available to the whole platform. In real life, the reach of this modification can be as great as the number of users in the networks; possibly millions.

However, simply increasing the total reach does not linearly augment the probabilities of altering

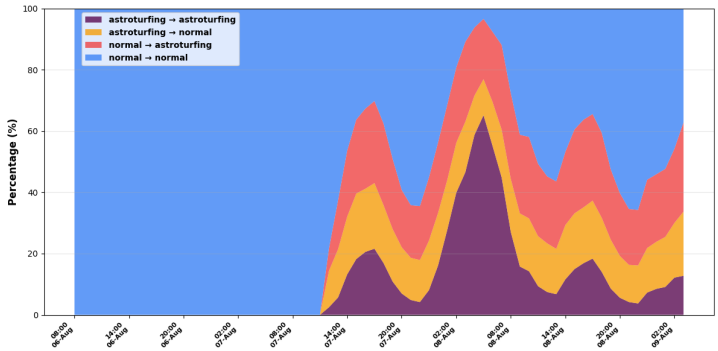
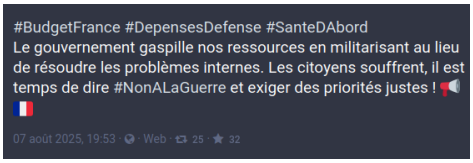


Figure 6: Percentage of cross-community interactions. We can observe that during the night, "normal" accounts interact mainly with the astroturfing accounts, as they are the most active group in the media at this hour. It constitutes an observation of an artificial echo chamber.

Period	Trending hashtags on Mastodon
Before the attack	#France, #Sécurité, #BudgetMilitaire, #Militaire, #Défense, #Budget, #SécuritéNationale, #FranceFirst, #DIY
After the attack	#France, #Sécurité, #BudgetMilitaire, #Militaire, #Défense, #Budget, #SécuritéNationale, #FranceFirst, **#NonALaGuerre**, **#PrioriteEducation**

(a) Trending hashtags on Mastodon, before and after the attack.



(b) Example of an adversarial message published during the simulation. The post received high exposition, with 25 retweets and 32 favourites.

Figure 7: Overview of Mastodon activity and adversarial messages during the simulation.

the target behaviour, reaching operational objectives (i.e., causing kinetic impacts). In fact, as reported in [25], even large-scale attacks can fail if they do not genuinely resonate with or effectively penetrate the target audience in a meaningful way. This specific part paves the way for further research on the cognitive aspect of this framework.

4.2. T0100.003 "Co-opt Influencers"

Our framework can be considered a sandbox for TTP simulation; in this section, we propose an overview of another TTP implementation, Co-opt Influencers. This TTP is deployed to concretize and raise awareness on the possible impacts of already corrupted influencers within online social networks, especially towards the youth.

4.2.1. Sources

This kind of attack is documented in Viginum Technical reports on interferences observed during Romania's 2025 presidential elections. "Co-opt Influencers", used along astroturfing techniques in the report, consists of using renowned influencers by corrupting them or using them as useful idiots to push a political agenda over their following base. This tactic is particularly effective as proved during Romania's Elections. Furthermore, it is regularly combined with Astroturfing to create "fake experts" or "fake influencers," that is, artificially amplified accounts designed to project crafted expertise on specific topics. By leveraging a perceived authority, these accounts can shape discourse, influence public opinion, and reinforce particular narratives in a way that appears authentic but actually is strategically orchestrated.

4.2.2. Technical Instantiation

For this part, the focus will be on how corrupted influencers can influence entire communities. To do so, the initial setup presented at Subsection 4.1 is restored, with the removal of the astroturfing communities in order to keep only the "normal" communities.

Among them, 14 (~2.6% of the accounts from the network) of the existing influencers are randomly selected (from Luminary, Celebrities, Opinion Leaders and Experts) to be the corrupted influencers, reproducing what happened in Romania's elections (based on Viginum technical report). Finally, their topic interest is set to privilege "Elections", aiming to share their point of view across the whole graph. For the entire topic distribution, the initialisation setup can be seen in Table 2.

Table 2
Initialisation of topic interests (mean +- std) by community group

Community Group	Climate	Elections	Foreign Policy	Immigration	Education Policy
Normal	0.23 +- 0.23	**0.15 +- 0.16**	0.15 +- 0.20	0.18 +- 0.21	0.13 +- 0.20
Corrupted influencers	0.0 +- 0.0	**1.0 +- 0.0**	0.07 +- 0.13	0.20 +- 0.20	0.01 +- 0.03

In order to track the reach of the messages pushed by the corrupted influencers, a first step is to monitor the co-retweet, co-favourites and replies network across the simulation, similarly as in [21]. Furthermore, it is essential to monitor the rate of change in interest levels to assess whether the topic is more and more relevant within the network. Should interest increase, the topic disseminated by the designated influencers is likely to acquire additional spreaders, thereby amplifying its overall reach.

Akin to the precedent astroturfing TTP, the corrupted influencers become active only after a 36 hours delay, i.e., the second day at noon.

4.2.3. Results

For this TTP, the goal is to observe how influencers, which are central in the graph, can influence the entire network towards specific topics. In real life, these influencers can be paid, used as useful idiots, or persuaded to spread propaganda over foreign states narratives.

With this limited number of influencers, Table 3 shows that shortly after the start of the attack, a large portion of the network starts to interact with them, highlighting their reach capabilities.

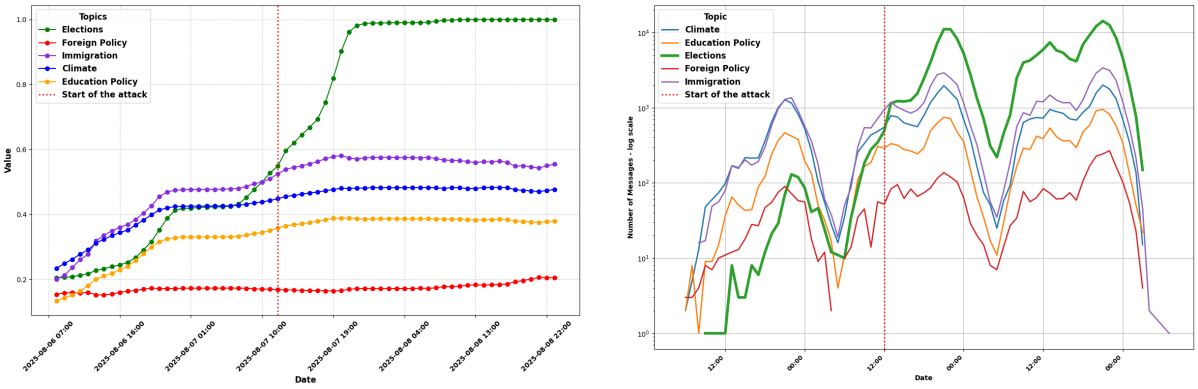
Table 3
Number of interactions to a content pushed by the corrupted influencers. Rapidly, a large proportion of normal nodes is interacting with the corrupted influencers, with a large volume of interactions. Note that the percentage is not cumulative, and is constrained over the requested period.

Period of the attack	12:00 to 21:00	21:00 to 6:00	6:00 to 15:00	15:00 to 24:00
Percentage of normal nodes interacting	68%	92.4%	89.9%	98.8%
Number of interactions	826	1275	1056	2219

Over the first day, almost all nodes in the other communities interacted at least once with the influencers' messages. This result means that these accounts can drive the chosen topics and make it trending over the platform, and give it a reach it did not have before. These interactions will drive up the interest rate of the specific chosen topic, which will cause additional engagement increase. Through Subfigure 8a, the average interest of the normal accounts grows way faster than the other interest rates after the start of the attack.

This increase in interest rate mechanically increases the number of interactions on these topics, as shown with Subfigure 8b. This subfigure also shows that while the number of interactions over Elections topic increases greatly after the start of the attack (the number of interaction mainly stays low before the attack), other topics are not impacted and stay at the same levels before and after the attack. This shows that the influencers drive the other communities over this particular topic, highlighting their impact.

These results show that a small number of influencers can influence entire communities, as highlighted in real life with Viginum's report about the Romanian elections. The cognitive process of narrative



(a) Rate of interest over topics across time. For the simulation, the threshold of activation is fixed at 0.60

(b) Number of interactions -posts, replies, favourites, retweets - (log scale) over topics across time.

Figure 8: Temporal evolution of interests and interactions during the simulation. The **Elections topic** interest rate starts at a low level and grows only linearly in the early phase. After the corrupted influencers launch their attack, interest in the topic surges dramatically, making it the dominant topic by a wide margin. The same pattern can be observed in the second figure, with low interaction amounts before the attack, and a great increase after the attack. The first 8 hours of the simulation are omitted to exclude the initialisation phase.

conversion was not modelled. In practice, however, influencers can persuade individuals to adopt their cause, potentially leading to kinetic decision-making.

5. Limitations

We showed in this paper that our simulation is credible with the objectives of training against reproduced TTPs. However, we did not take into account all the social effects within the social networks.

Namely, for the decision to follow and unfollow, we did not introduce polarity over topics. In fact, one can be really interested in a topic but in an opposite way of others. This two-faced coin can be primordial in reproducing coherent follow updates (as aligned users tend to follow themselves more frequently than unaligned ones). Similarly, evolution in topic interest is also based on more socioeconomic variables, as well as cultural characteristics from individuals.

Furthermore, we simulated only five topics over a three days simulation, but real social networks contain way more parallel topics of discussion. Cross-interest between topics is also way higher in the real life, with topics impacting each others massively. We illustrated this capacity with light drops of interest on the second TTP, especially with the topic Immigration.

Moreover, reproducing credible full-scale attacks is far from trivial, as defenders do not have the full overview over what the attackers did, and especially what they wanted to do.

Finally, we did not evaluate the quality of the text generation in this work, but text is very important for conveying information in a specific manner. Attackers understand that importance, and play on sentiments such as fear or surprise to elicit reactions from users, maximising the impact of their campaigns.

6. Conclusion

In this paper, we presented a system used to simulate credible social networks, used to train specialised analysts against influence operations.

These analysts require the simulation of specific Tactics, Techniques, and Procedures (TTPs) to capitalise on existing knowledge and mitigate their impact. Enabling this simulation, we presented in this work a system generating realistic dynamic and controllable social graphs, allowing an end-user to parametrise TTPs to be reproduced. The system encompasses three distinct modules: the initial community generation, the community linking, and the information dissemination modules. The first module

generates user accounts (nodes in a graph) of different types and their follower distribution, according to end-user parameters. The second module serves the purpose of defining how communities are linked, influencing how the simulation will unfold. Finally, the last module aims to simulate information diffusion in a graph, adding multiple variables tailored to the specific needs of training against influence operations.

To illustrate the coherence of the simulation, we have compiled two distinct case studies. The first one simulated an astroturfing campaign that involved two bot-heavy communities attacking two normal communities. The second example is used to illustrate and analyse how paid influencers can influence entire communities, similar to COVID-19 pandemic disinformation campaigns. Through our results, we show that our simulations are coherent with the literature observations of reported attacks, and that the simulation is credible with respect to real platforms. This credibility allows trainers to use this system to enhance the trainees' immersion and thus, the training effect.

Declaration on Generative AI

During the preparation of this work, the authors used Mistral in order to perform grammar and spelling checks. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. Morteo, TO CLARIFY THE TYPIFICATION OF INFLUENCERS: A REVIEW OF THE LITERATURE, 2018.

[2] M. Mazza, M. Avvenuti, S. Cresci, M. Tesconi, Investigating the difference between trolls, social bots, and humans on Twitter, Computer Communications 196 (2022) 23–36. URL: https://www.sciencedirect.com/science/article/pii/S0140366422003711. doi:10.1016/j.comcom.2022.09.022.

[3] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, E.-P. Lim, On Profiling Bots in Social Media, in: Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2016, pp. 92–109. URL: https://doi.org/10.1007/978-3-319-47880-7_6. doi:10.1007/978-3-319-47880-7_6.

[4] L. G. Mojica, Modeling Trolling in Social Media Conversations, 2016. URL: http://arxiv.org/abs/1612.05310, issue: arXiv:1612.05310 arXiv:1612.05310 [cs].

[5] D. L. Linvill, , P. L. Warren, Troll Factories: Manufacturing Specialized Disinformation on Twitter, Political Communication 37 (2020) 447–467. URL: https://doi.org/10.1080/10584609.2020.1718257. doi:10.1080/10584609.2020.1718257, publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1718257.

[6] A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512. URL: http://arxiv.org/abs/cond-mat/9910332. doi:10.1126/science.286.5439.509, arXiv:cond-mat/9910332.

[7] M. Li, X. Wang, S. Zhang, A Survey on Information Diffusion in Online Social Networks: Models and Methods, Information 8 (2017). doi:10.3390/info8040118.

[8] D. Kempe, J. Kleinberg, E. Tardos, Influential Nodes in a Diffusion Model for Social Networks, in: D. Hutchison, T. Kanade, J. Kittler, a. et (Eds.), Automata, Languages and Programming, volume 3580, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 1127–1138. URL: http://link.springer.com/10.1007/11523468_91. doi:10.1007/11523468_91, series Title: Lecture Notes in Computer Science.

[9] N. Barbieri, F. Bonchi, G. Manco, Topic-Aware Social Influence Propagation Models, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, Brussels, Belgium, 2012, pp. 81–90. URL: http://ieeexplore.ieee.org/document/6413913/. doi:10.1109/ICDM.2012.122.

[10] C. Liu, Z.-K. Zhang, Information spreading on dynamic social networks, Communications in

Nonlinear Science and Numerical Simulation 19 (2014) 896–904. URL: https://www.sciencedirect.com/science/article/pii/S100757041300378X. doi:10.1016/j.cnsns.2013.08.028.

[11] R. Ross, An application of the theory of probabilities to the study of a priori pathometry.—Part I, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 92 (1916) 204–230. URL: https://royalsocietypublishing.org/doi/10.1098/rspa.1916.0007. doi:10.1098/rspa.1916.0007.

[12] S. Kumar, M. Saini, M. Goel, B. S. Panda, Modeling information diffusion in online social networks using a modified forest-fire model, Journal of Intelligent Information Systems 56 (2021) 355–377. URL: https://link.springer.com/10.1007/s10844-020-00623-8. doi:10.1007/s10844-020-00623-8.

[13] B. Nimmo, Anatomy of an Info-War: How Russia's Propaganda Machine Works, and How to Counter It, 2015. URL: https://www.stopfake.org/en/anatomy-of-an-info-war-how-russia-s-propaganda-machine-works-and-how-to-counter-it/, section: Context.

[14] C. François, Actors, Behaviors, Content: A Disinformation ABC (????).

[15] S. Blazek, SCOTCH: a framework for rapidly assessing influence operations, 2021.

[16] K. M. Carley, Social cybersecurity: an emerging science, Computational and Mathematical Organization Theory 26 (2020) 365–381. URL: https://doi.org/10.1007/s10588-020-09322-9. doi:10.1007/s10588-020-09322-9, number: 4.

[17] D. Schoch, F. B. Keller, S. Stier, J. Yang, Coordination patterns reveal online political astroturfing across the world, Scientific Reports 12 (2022) 4572. URL: https://www.nature.com/articles/s41598-022-08404-9. doi:10.1038/s41598-022-08404-9, publisher: Nature Publishing Group.

[18] L. Vargas, P. Emami, P. Traynor, On the Detection of Disinformation Campaign Activity with Network Analysis, in: Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, ACM, Virtual Event USA, 2020, pp. 133–146. URL: https://dl.acm.org/doi/10.1145/3411495.3421363. doi:10.1145/3411495.3421363.

[19] M. Fernandez, A. Bellogín, I. Cantador, Analysing the Effect of Recommendation Algorithms on the Spread of Misinformation, in: ACM Web Science Conference, ACM, Stuttgart Germany, 2024, pp. 159–169. URL: https://dl.acm.org/doi/10.1145/3614419.3644003. doi:10.1145/3614419.3644003.

[20] G. Gadek, Détection d'opinions, d'acteurs-clés et de communautés thématiques dans les médias sociaux, phdthesis, Normandie Université, 2018. URL: https://theses.hal.science/tel-02064171.

[21] PROMPT / Narrative Intelligence for Information Integrity / PROMPT, 2025. URL: https://disinfo-prompt.eu/.

[22] L. Manikonda, G. Beigi, H. Liu, S. Kambhampati, (PDF) Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo through the Lens of Social Media, 2018. URL: https://www.researchgate.net/publication/323931993_Twitter_for_Sparking_a_Movement_Reddit_for_Sharing_the_Moment_metoo_through_the_Lens_of_Social_Media. doi:10.48550/arXiv.1803.08022.

[23] J. Schler, E. Bonchek-Dokow, Profiling Astroturfers on Facebook: A Complete Framework for Labeling, Feature Extraction, and Classification, Machine Learning and Knowledge Extraction 6 (2024) 2183–2200. URL: https://www.mdpi.com/2504-4990/6/4/108. doi:10.3390/make6040108, publisher: Multidisciplinary Digital Publishing Institute.

[24] U. Oliveri, G. Gadek, A. Dey, B. Costé, D. Lolive, A. Delhay-Lorrain, B. Grilheres, SocialForge: Simulating the Social Internet to Provide Realistic Training Against Influence Operations, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics 2025, Industry Track, Association for Computational Linguistics, Vienna, Austria, 2025.

[25] G. Eady, T. Paskhalis, J. Zilinsky, R. Bonneau, J. Nagler, J. A. Tucker, Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior, Nature Communications 14 (2023) 62. URL: https://www.nature.com/articles/s41467-022-35576-9. doi:10.1038/s41467-022-35576-9, publisher: Nature Publishing Group.