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

Ulysse Oliveri, Alexandre Dey, Guillaume Gadek

DEFENCE AND SPACE / IRISA Expression

C&ESAR 2025 by DGA
Thursday, November 20, 2025
Rennes, France

AIRBUS

Introduction

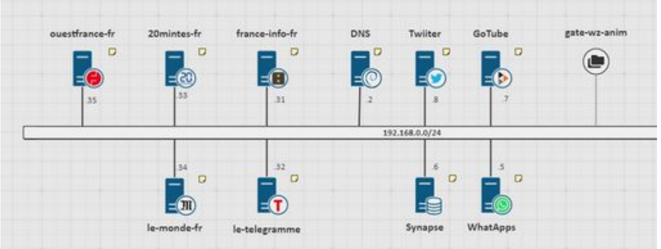
- Modern influence operations on social media are **complex** and **high velocity.**

- Analysts reinforce their capacities with **continuous training**, **updating** their **methodology** and their **knowledge** on new attacks methods.

- **Online behaviors** and their **associated interactions** are **produced** within a **closed and controllable infosphere**

Source: Viginum's Matriochka report

Source: Armée de Terre - Defnet 2025

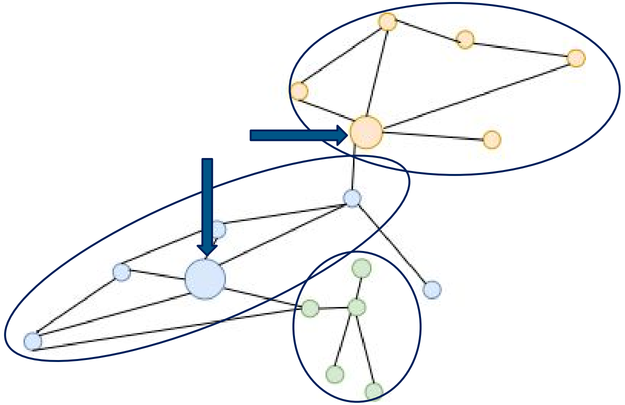


Airbus Cyberrange

AIRBUS

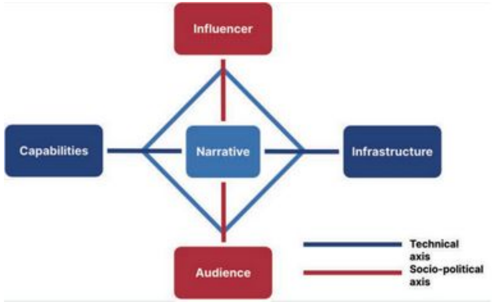
Training objectives

Informational environment cartography

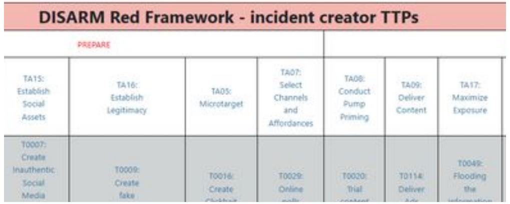


Simulate social graph topology and information diffusion

Inauthentic or coordinated behaviors detection



Source: Diamond model for influence - RecordedFuture



DISARM Red Framework - incident creator TTPs

PREPARE						
TA15: Establish Social Assets	TA16: Establish Legitimacy	TA05: Microtarget	TA07: Select Channels and Affordances	TA08: Conduct Pump Priming	TA09: Deliver Content	TA17: Maximize Exposure
T0007: Create Inauthentic Social Media	T0009: Create fake	T0016: Create	T0029: Online	T0020: Trial	T0114: Deliver	T0049: Flooding the

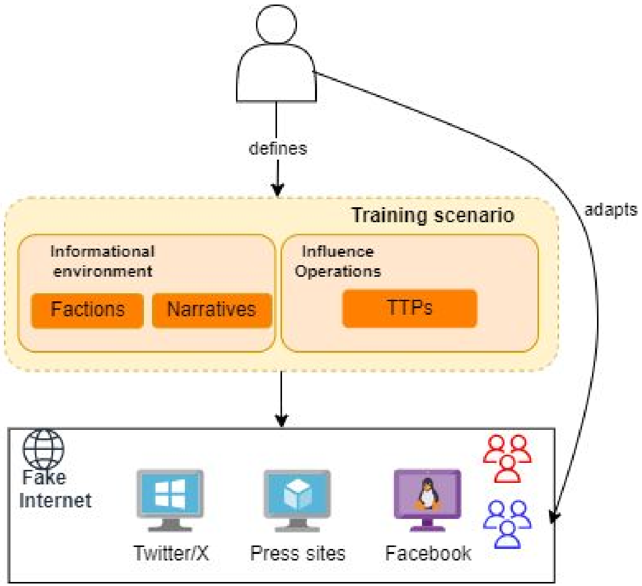
Enable the simulation of coherent TTPs

Engagement action

Source: RFI

Integrate the players into the simulation

Training unfolding



DISARM Red Framework - incident creator TTPs																
PLAN			PREPARE						EXECUTE							ASSESS
TA01: Plan Strategy	TA02: Plan Objectives	TA13: Target Audience Analysis	TA14: Develop Narratives	TA06: Develop Content	TA15: Establish Social Assets	TA16: Establish Legitimacy	TA05: Microtarget	TA07: Select Channels and Affordances	TA08: Conduct Pump Priming	TA09: Deliver Content	TA17: Maximize Exposure	TA18: Drive Online Harms	TA10: Drive Offline Activity	TA11: Persist in the Information Environment	TA12: Assess Effectiveness
T0073: Determine Target Audiences	T0002: Facilitate State Propaganda	T0072: Segment Audiences	T0003: Leverage Existing Narratives	T0015: Create hashtags and search artifacts	T0007: Create Inauthentic Social Media Pages and Groups	T0009: Create fake experts	T0016: Create Clickbait	T0029: Online polls	T0020: Trial content	T0114: Deliver Ads	T0049: Flooding the Information Space	T0047: Censor social media as a political force	T0017: Conduct fundraising	T0059: Play the long game	T0132: Measure Performance

Source: Disarm Explorer

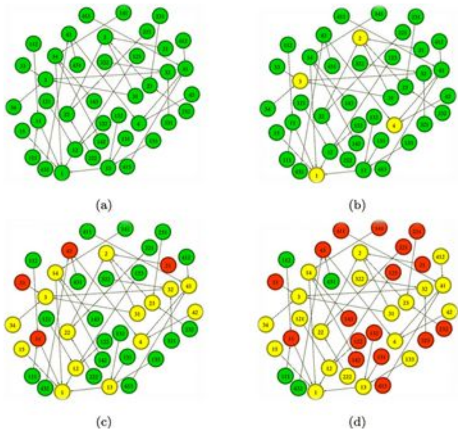
AIRBUS

Related Work

- The **users differ** in:
  - Their **connectivity in the social graph -** *Morteo et al 2018*
  - **Post and reaction patterns** - *Mazza et al 2022*
  - **Temporal interaction** patterns - *Oentaryo et al 2016*
  - The **intents** behind the interactions - *Mojica et al 2016*

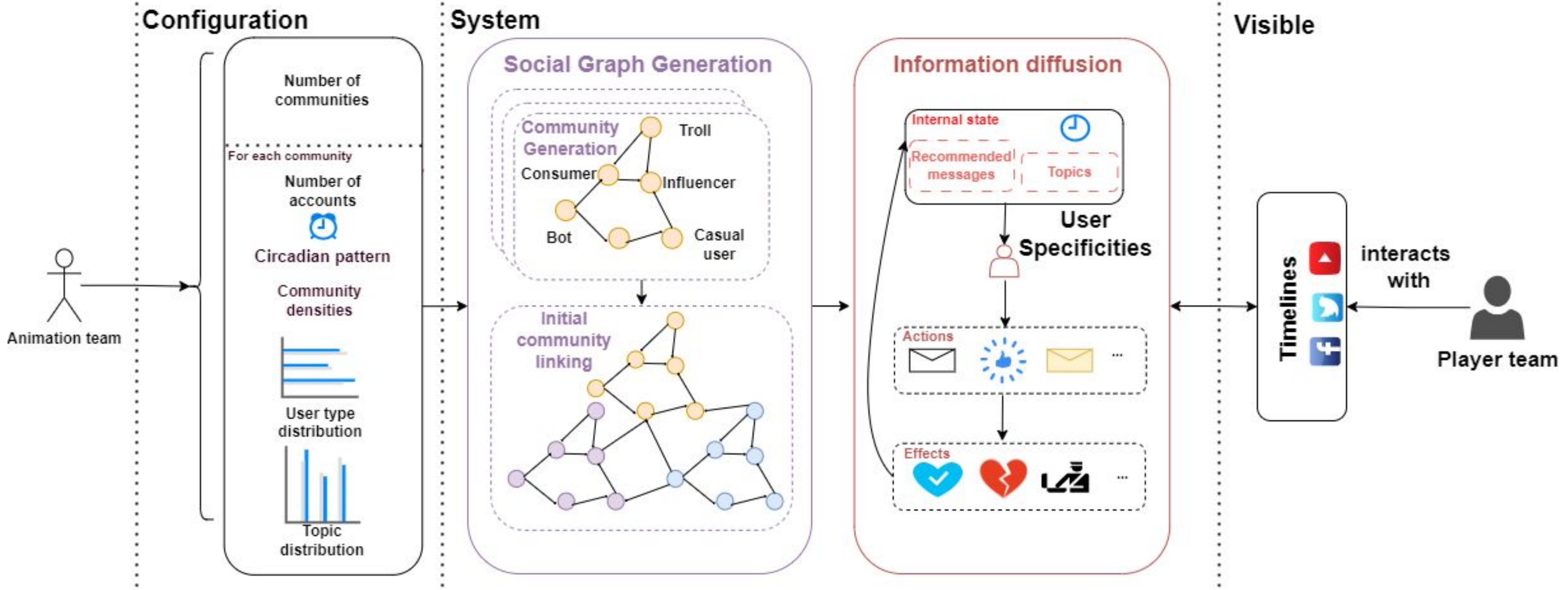
- From these characteristics emerge these type of users:
  - **Influencers** (Luminary, celebrities, expert, etc.) - *Morteo et al 2018*
  - **Casual users or Consumers** (a.k.a lurkers and networkers) - *Morrison et al 2013*
  - **Trolls** (Left-troll, Right-troll, etc.) - *Linvill et al 2020*
  - **Bots** (Spam Bots, Content Bots, Engagement Bots) - *Oentaryo et al 2016*

- Influence operations **manipulate** the social networks **recommendation systems**:
  - Using **specific keywords -** *Viginum's report on Romania's elections*
  - Promoting **divisive topics -** *Zhang et al 2017*
- To model the user accounts, their interactions, and the adversarial behaviors we inspire from **Modified Forest-Fire Algorithm** (*Kumar et al 2021)*



Modified Forest-fire model - Kumar et al 2021

Social network population



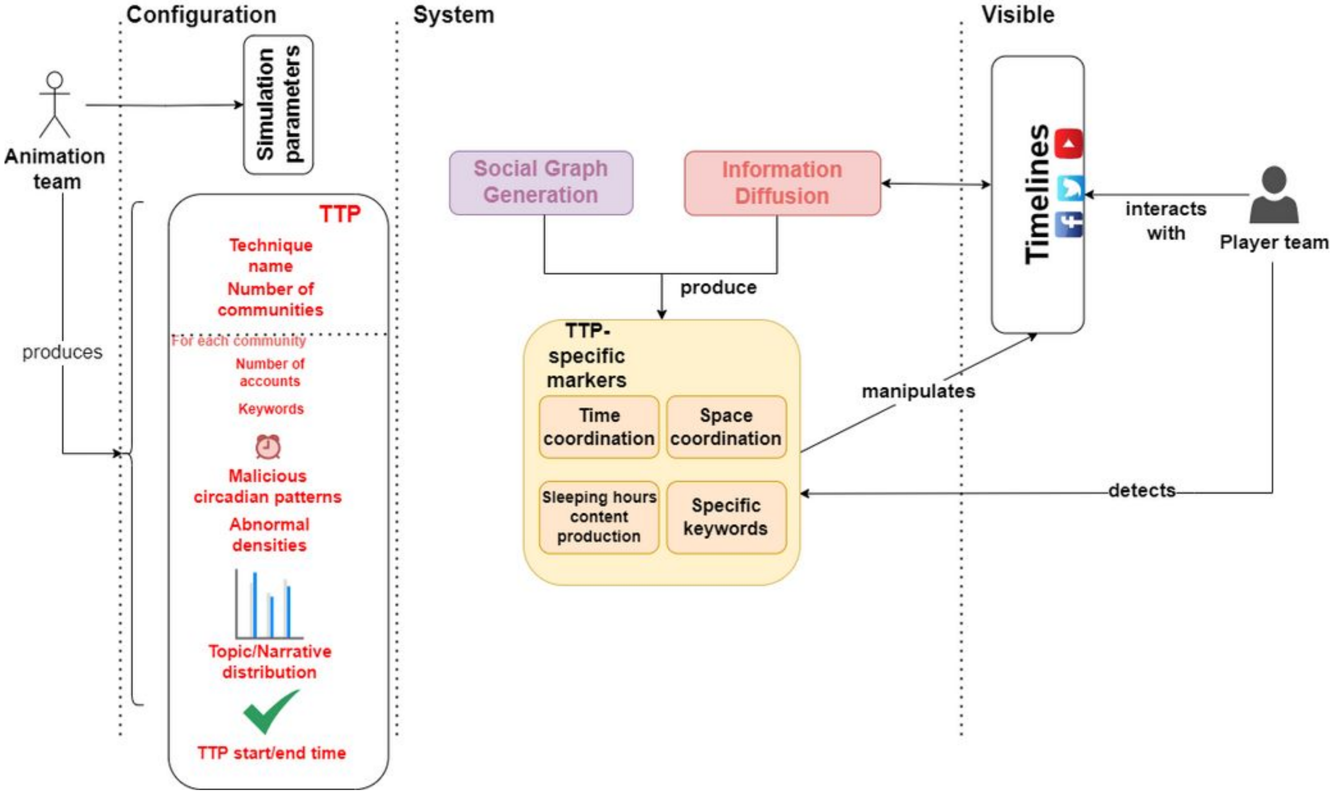
Enables user-tailored community generations

Simulates credible information diffusion within social networks

Allows the population of simulated social network

AIRBUS

Enabling the TTP simulation



The focus is on DISARM tactics that alter the social graphs topology and manipulate the information space.

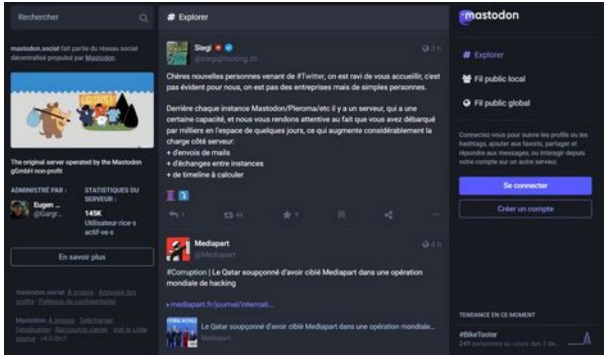
The system modules should **produce TTP-specific markers**

The final aim is that the **trainees detect these TTPs** using its usual procedures

Introduction to cases studies

Informational environment

- Simulating Twitter/X with a local-hosted **Mastodon**
- Generating the contents using our **own data generation system** (SocialForge)



Emulated adversary

Technique name	Description	Desired effects
T0099.001: Astroturfing	Synthetic consensus over specific topics, occupying the information space	- Push network-wide narratives - Content spikes during off-hours - Driving the dynamism of the platform
T0100.003 Co-opt Influencers:	Using renowned influencers by corrupting them or using them as useful idiots to push a political agenda over their following base	- Push network-wide narratives - Drive the interest for a specific topic

Case study 1

T0099.001: Astroturfing

Astroturfing - Initial Setup

- We generate **520 genuine** and **131 malicious accounts**
- **Astroturfing communities** contain an **higher proportion of bots and trolls** w.r.t genuine accounts
- We give a set of **ten keywords** to the astroturfers to **employ in their contents**
- The system **starts** the attack **after one day and a half of simulation**

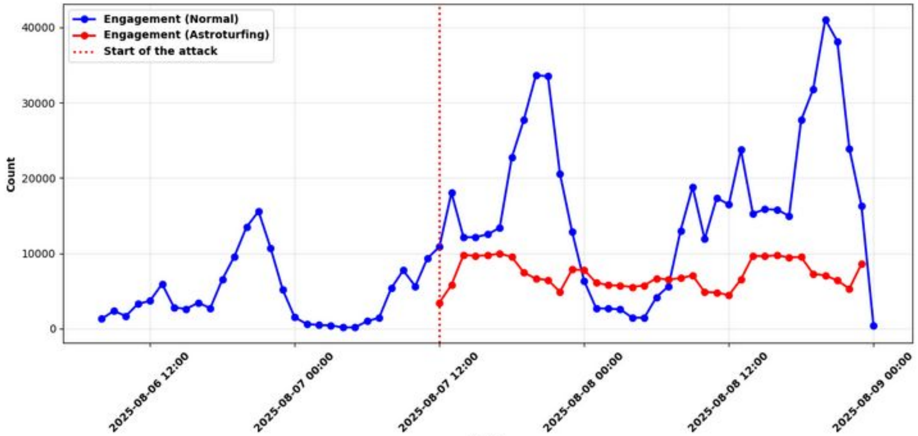
	Intra-community density	**Interaction rhythm**	**Prevalent account types**
Malicious accounts	0.30	Mainly off-activity hours	Trolls, Bots
Genuine accounts	0.10	Circadian rhythm	Consumers, Casual users

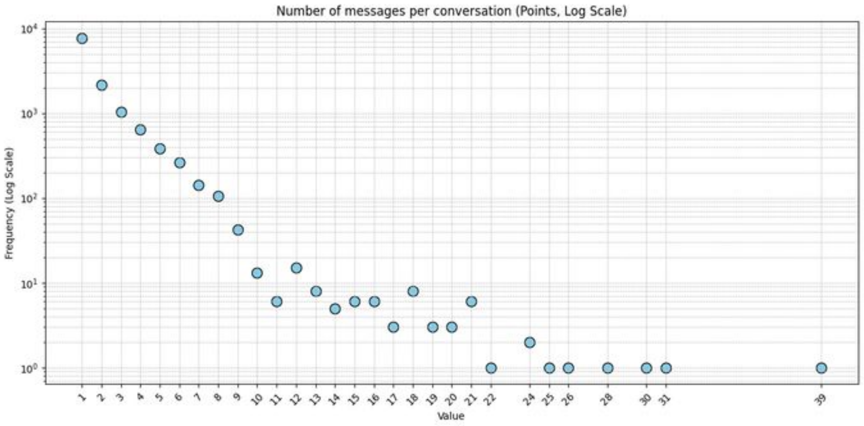
Astroturfing - Evaluation Setup

Social networks metrics	**Simulation evolution metrics**	**Narrative exposure**
- Distribution of the engagement - Activity circadian patterns	- Evolution of the follows / unfollows - Community distribution	- Cross-community interactions distribution - Mastodon trendings
Evaluates the **coherence / credibility** of the interactions	Evaluates the **dynamism** of the **simulation**	Evaluates the **visibility** of the **attack**

engagement = Number of tweets + replies + favorites + retweets

Astroturfing - Social Network metrics





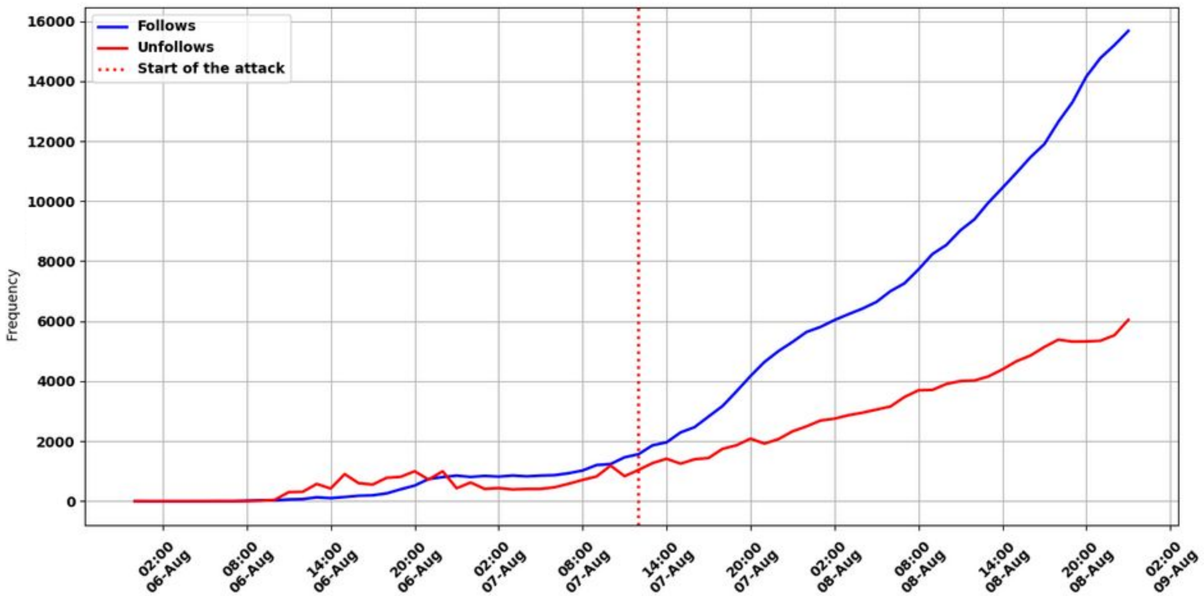
The engagement of the **genuine community aligns with circadian patterns**

Astroturfing communities **peak during off-hour, as observed in the literature**

The **number of engagement per conversation** also **follows a heavy-tail distribution**, **as observed in common social network metrics**

Astroturfing - Simulation evolution

Modularity score



Before the attack: **0.6**

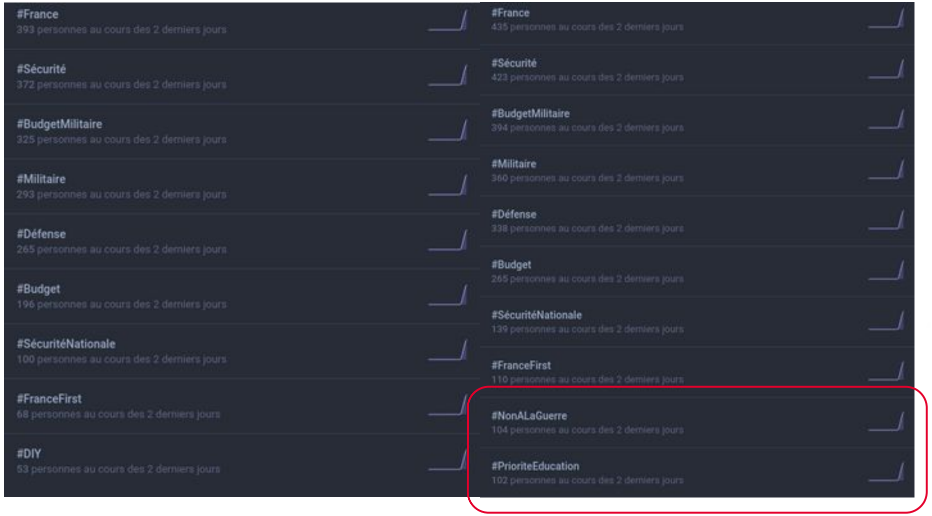
- The follows and unfollows are **stable**
- **High modularity score**, indicating distinct communities

After the attack: **0.2**

- **Drastic increase** with the astroturfing following the normal communities.
- **Lower modularity score**, communities are merging

Astroturfing - Narrative exposure

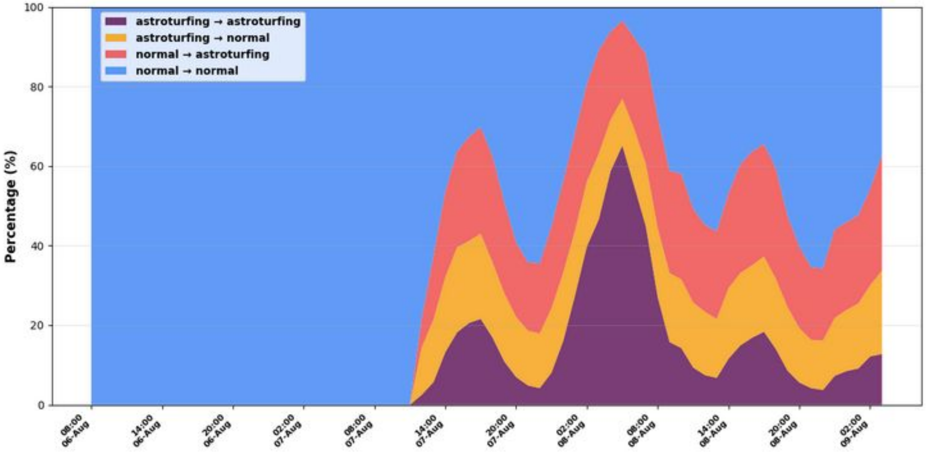
Mastodon trendings before and after the attack



Before the attack

After the attack

Cross-community interactions ratio between astroturfing accounts and genuine communities



Astroturfers are implicated in ~50% interactions over the platform despite the 1 to 5 prevalence ratio.

Appearance of adversarial hashtags in the Mastodon trendings which exposes it to the whole platform

The astroturfing accounts monopolize the attention and give a synthetic sentiment of consensus within the platform

Case study 2
T0100.003 "Co-opt Influencers"

Co-opt Influencers - Initial Setup

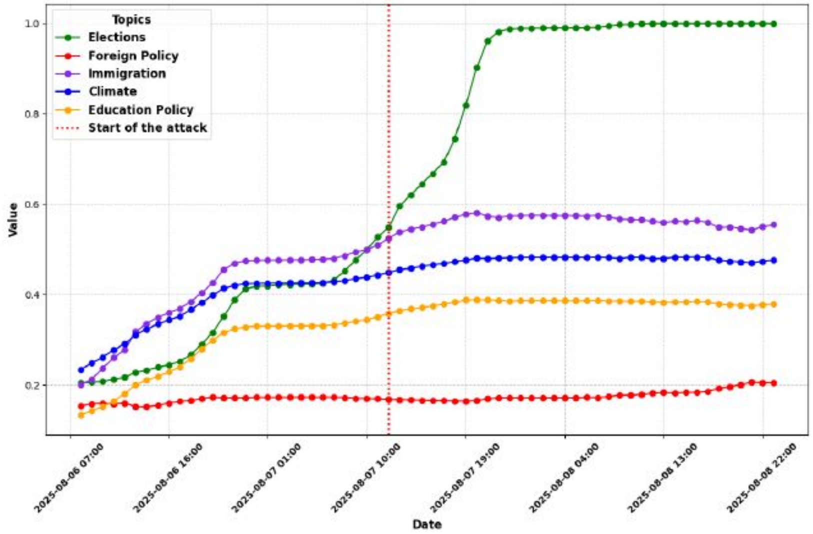
- **520** genuine **accounts.**
- **14** (**~2.6%** of the accounts from the network) **influencer** accounts **are corrupted.**
- **The corrupted** accounts **highly privilege the Elections** topic

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

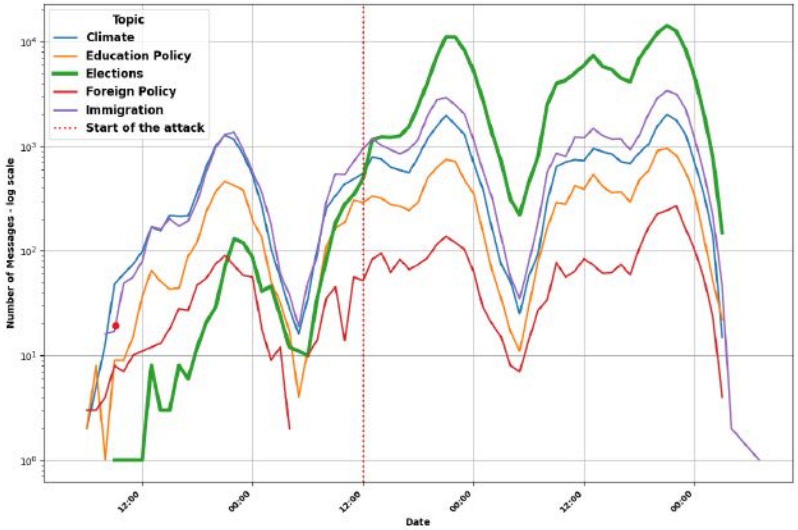
Co-opt Influencers - Evaluation Setup

Simulation evolution metrics	**Narrative exposure**
- Topic interest over the simulation - Monitoring the engagement across topics	- Accounts visibility in the platform
Evaluates the topic engagement and interest rate changer over of the **simulation**	Evaluates the **visibility** of the **attack**

Co-opt Influencers - Simulation evolution metrics



(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.

The Election topic spikes both in interest and in engagement shortly after the beginning of the attack

Co-opt Influencers - Narrative exposure

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

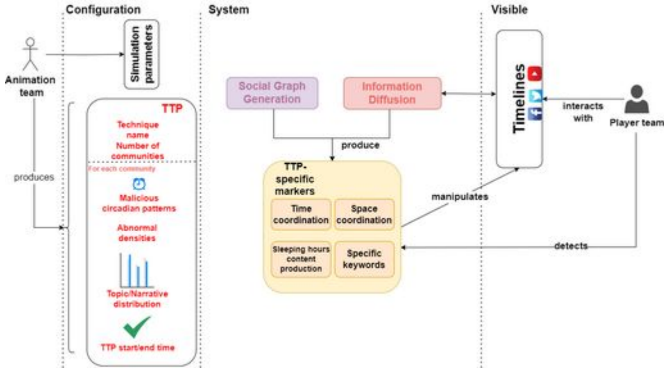
The influencers new narrative becomes highly discussed within the network

Almost all of the network has interacted with the new narrative at the end of the simulation

Conclusion and Perspectives

Achievements:

- We presented a **framework to simulate adversarial TTPs** for training exercises
- We implemented **two specific TTPs**; **astroturfing operations and corrupted influencers**
- We assessed this framework through **topological** and **information diffusion measures**
- We deployed a **self-hosted Mastodon** to **generate the diverse interactions and the trendings modifications**



Perspectives:

- **Persistence** over the long term of the simulated behaviors
  - **Decomposition** of the infops in **several intermediary steps** (e.g., step 1: infiltrate the network, step 2: test the defences, etc.)
  - **Coordination** of distinct TTPs
- **Study behavioral changes** across time (i.e., genuine accounts get converted to adversarial narratives)
- **Evaluate the contents semantics** between the diverse user types

More details on the data generator

SocialForge, Oliveri et al

Airbus Influence warfare training platform

Airbus L2I

AIRBUS