Evaluating Policy Synergies Against Online Extremism

An Agent-Based Simulation and Cost-Benefit Analysis

Matteo Vergani, Andrea Giovannetti, Stephanie Zi Xin Ng, Chee Peng Lim, James Zhang, Robin Scott

Presented by:

Andrea Giovannetti

Peter Faber Business School, Australian Catholic University

Violence Research Centre, University of Cambridge

ag2051@cam.ac.uk

Download Link

Contact

Motivation

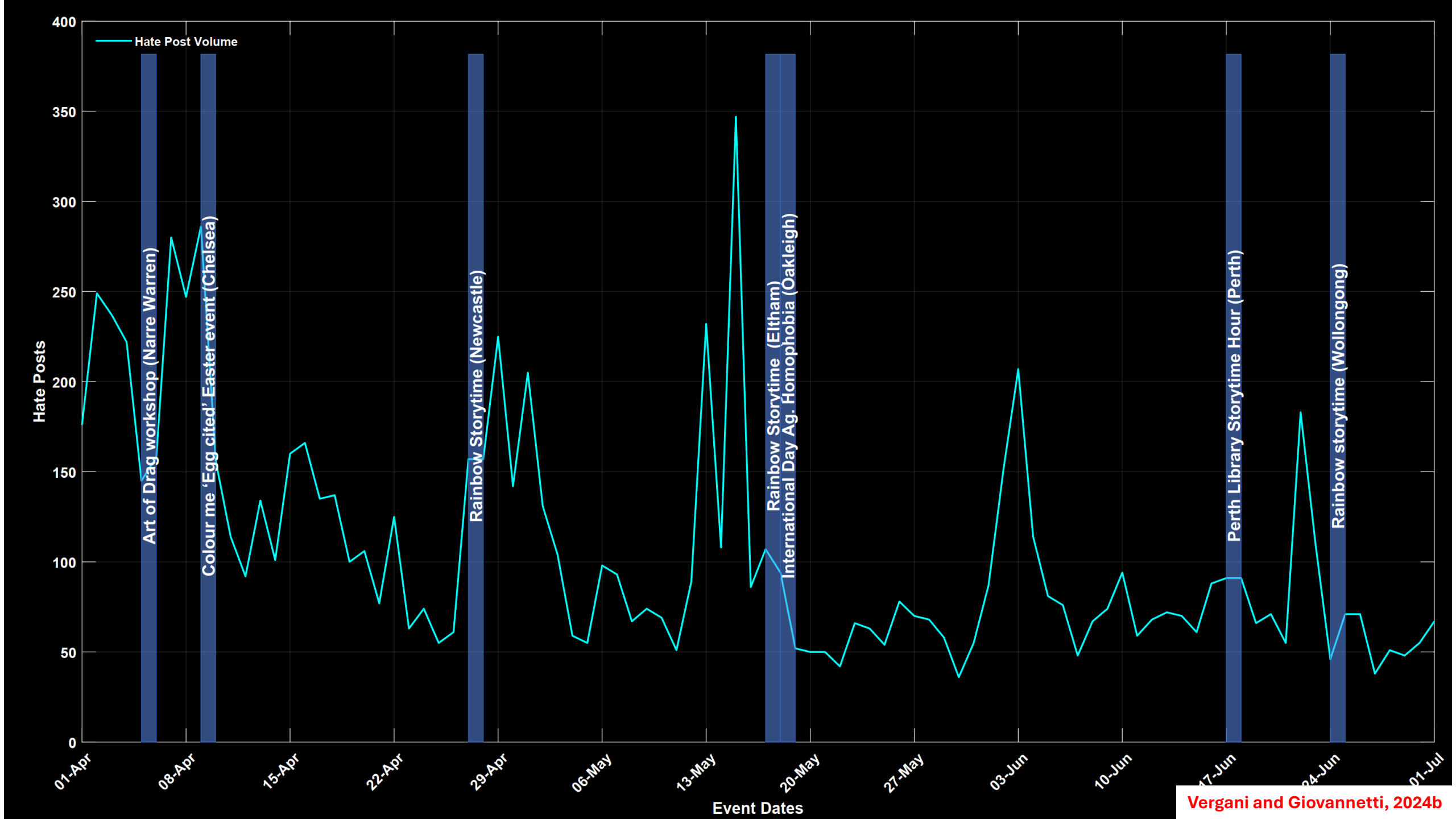
- **Online platforms** are a key **channel** for the spread of **online hate and extremism**, globally

- Evidence is building up on **transmission chains** between **online hate** and **offline incidents.**

- Online hate can be **weaponized** by **foreign states and state-like actors** to destabilize democracies and put the social fabric at risk.

Example:

Quran Burning Rallies in Sweden (July 2023)



Vergani and Giovannetti, 2024b

Motivation

(Working) Definition of Extremism:

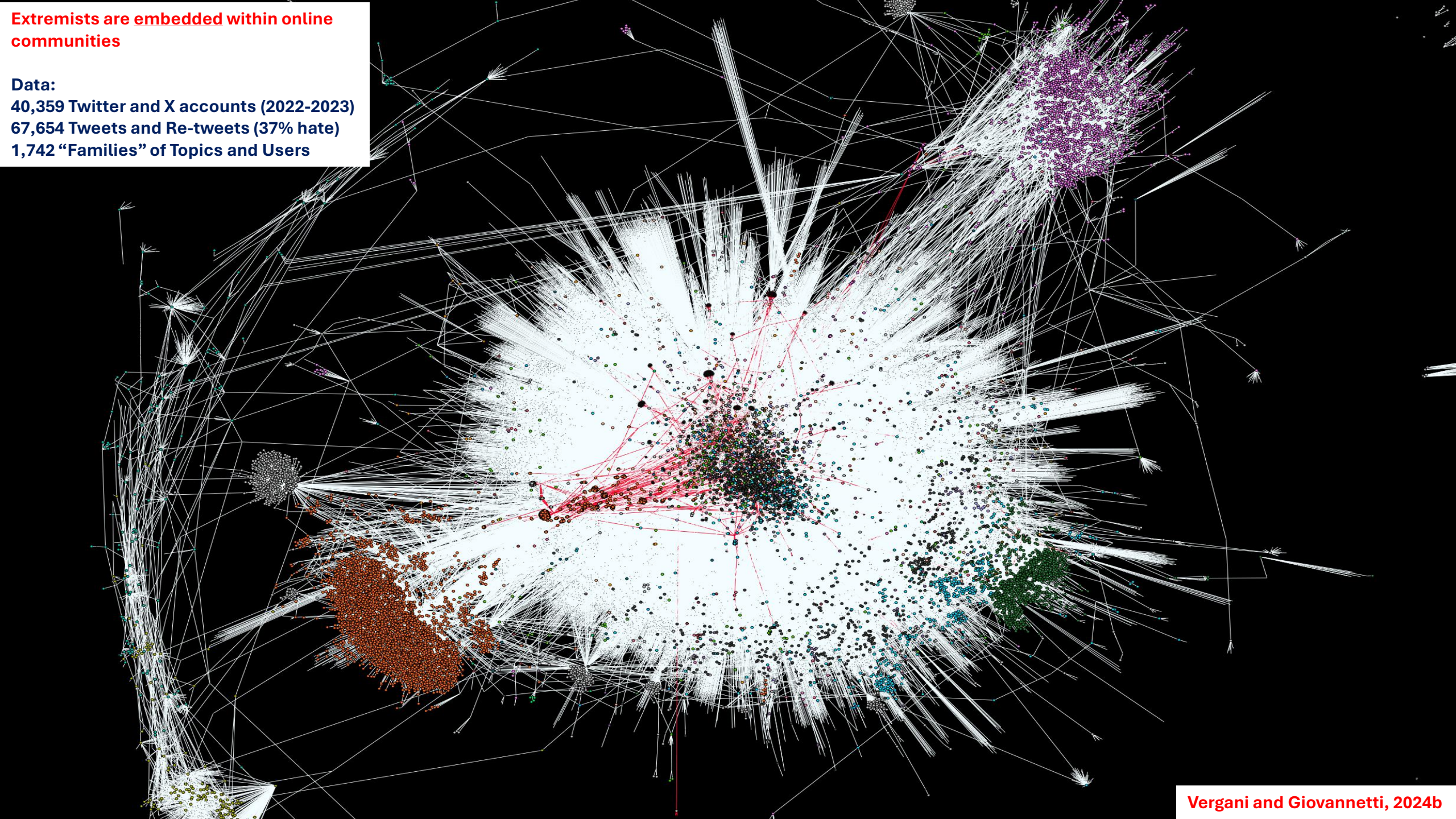
The beliefs and actions of people who support or use violence to achieve ideological, religious or political goals. (Australian Government)

Extremist content **influences** the views of those who share salient interests, demographic features, ideological leanings with extremists.

Micro-dynamics are key for the diffusion chain to work

Extremists are embedded within online communities

Data:
40,359 Twitter and X accounts (2022-2023)
67,654 Tweets and Re-tweets (37% hate)
1,742 "Families" of Topics and Users



Vergani and Giovannetti, 2024b

Vergani and Giovannetti, 2024b

Motivation

Repressive counter-extremism strategies (e.g. **monitoring performed by human moderators and & deplatforming**) of extremists **is expensive, inefficient, socially costly and can backfire.**

An Alternative Approach

Soft Measures align better with democratic principles by promoting engagement, empowerment, and resilience within communities.

Examples: education, training, **inoculation programs**

Problem: Soft Measures are **hard to measure (hence, to price!)**

In 2019, Allen Consulting analyzed a suite of 13 Community-based Countering Violent Extremism programs activated in NSW (Australia) between 2015-2019 for a total cost of $47 AUD million and a reach of about 1.5 million inhabitants:

1. *There are challenges in measuring and quantifying outcomes of CVE, which can be attributed to […] what indicators should be tracked*
2. *About a third of programs failed to provide quantitative evidence on the capability of building societal protective factors against VE*

An Alternative Approach

Inoculation:

By exposing individuals to weakened forms of an argument and refuting it, individuals develop resistance to the argument

Mode of action

(1) Passive Inoculation: present and disprove an argument

(2) Active Inoculation: involve the subject in the process of refutal

Problem

- We know that inoculation works **at individual level**
- Example: in a review of 40 studies involving 10,000 participants, Banas and Rains (2010) found a direct effect
- However, results are from small samples and in isolation from other approaches.

More in general, it is hard to evaluate both families of policies

- **How do we measure/compare the functioning and the value of counter-extremism measures?**
- Gathering reliable data on the effectiveness of measures is challenging:
  1. **Ethical concerns** around designing randomised controlled trials.
  2. **Limited access** to proprietary data from social media platforms.
  3. **Reluctance** of users with extremist views to participate in studies.

Our Approach

- We generate data through an **agent-based modelling** (ABM) approach to overcome these challenges.
- An agent-based model is a **simulation** of the reality

- The ABM is **generalizable, replicable**, **falsifiable** and is **calibrated** with **evidence-based parameters**

- We use the ABM to assess **2 policies**:
  - Repressive policy: **deplatforming** of extremist users
  - Soft policy: **inoculation**

Scenario

- We generate a **social network** with topological features similar to a **prototypical anti-immigration Facebook group**. Individuals are heterogeneous and realistic.

- **Parameters** are calibrated using data from systematic reviews and empirical studies conducted in the field of terrorism research.

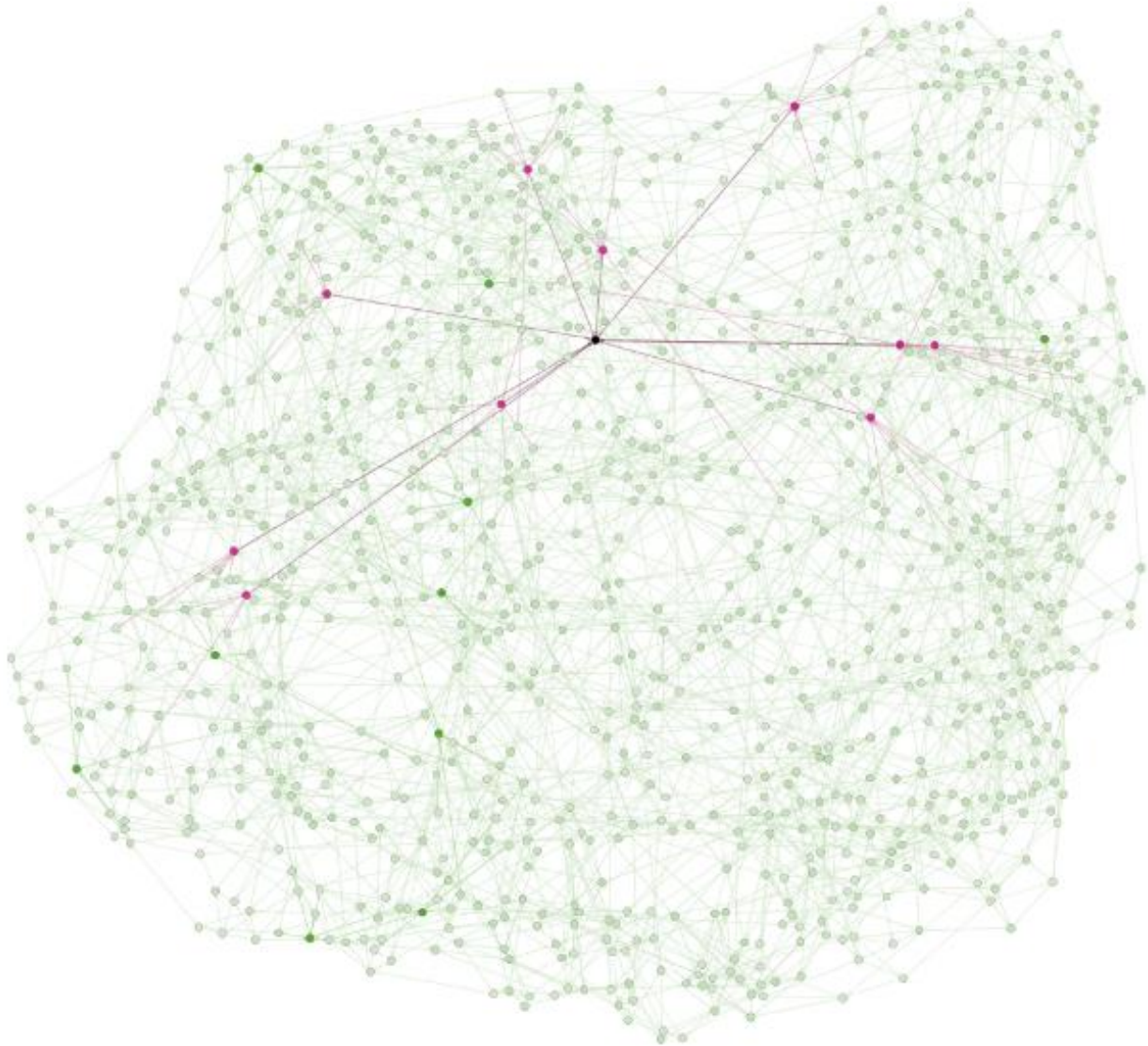
- In particular:
  - **Inoculation effect at individual level**
  - **real world cost figures**

- **Important:** This method allows us to assess policies **in isolation and in combination!**

Agent roles and interactions

(1) Extremists:

Spread radical content to users they are linked to



Agent roles and interactions

(2) Moderators:

Monitors and moderates **online** activities, identifying and deplatforming extremists.

The moderator is “**inefficient”**

- Pools a **subset** of the population
- Assesses a number of posts produced by the user in the previous years
- Decides whether to deplatform or not
- **This operation takes time**!

Agent roles and interactions

(3) Users:

- Members of a Facebook group, individually varying in characteristics (e.g. ideology, demography, education level, connections)
- Age, gender and connectivity is calibrated on META user profiles (Ramshaw, 2024)
- Additional attributes (e.g. education, marital status, criminal history) are obtained from ABS tables
- Attributes determine the **Susceptibility** to extremism narrative through the effect estimations of Wolfowicz et al. (2020)

Attribute	Factor Type	Effect Size	Odds Ratio	Definition
Age	Protective	-0.053	0.825	1 if age $>$ 35; 0 otherwise
Education	Protective	-0.039	0.868	1 if bachelor's degree or above; 0 otherwise
Marital Status	Protective	-0.038	0.871	1 if married; 0 otherwise
Gender	Risk	0.082	1.347	1 if male; 0 if female
Employment Status	Risk	0.042	1.165	1 if unemployed or no study; 0 otherwise
Criminal History	Risk	0.331	3.397	1 if committed crime; 0 otherwise

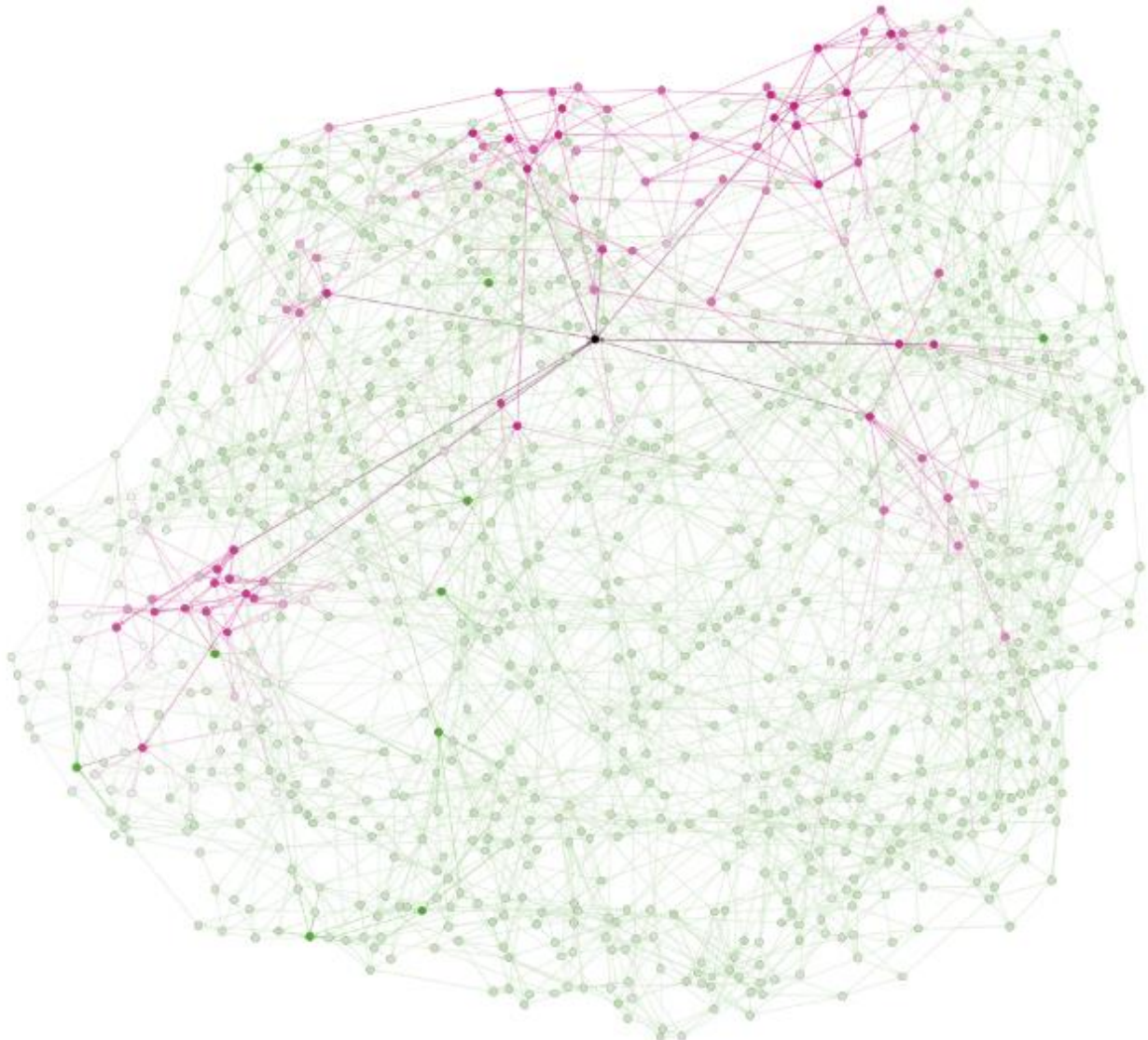
Table 5: Demographic factors used to calibrate users' susceptibility to extremist content.

Agent roles and interactions

Users perform two actions:

1) Users adjust their belief on extremist content conditional on:
   - Friends' Posts
   - Direct messages exchanged with the extremists they're linked to
   - Inoculators' activity

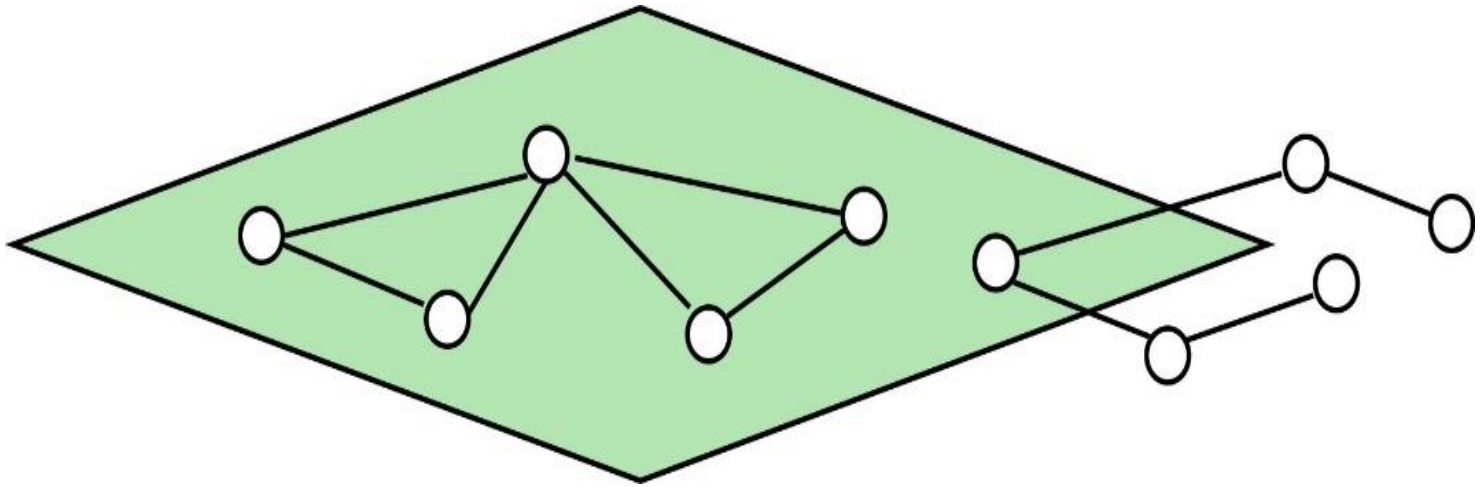
2) **If radicalized:** users will pass radical content to their contacts



Agent roles and interactions

(4) Inoculation Policy:

Deliver **offline** educational interventions, aiming to build resistance to extremist views.



Data

We **generate** data by manipulating **two** dimensions simultaneously:

(A) Radius of Moderator.

- From **1 ‰ to 1%** of population monitored every 10 days [11 possible levels of moderation]

(B) Frequency of Inoculation Events.

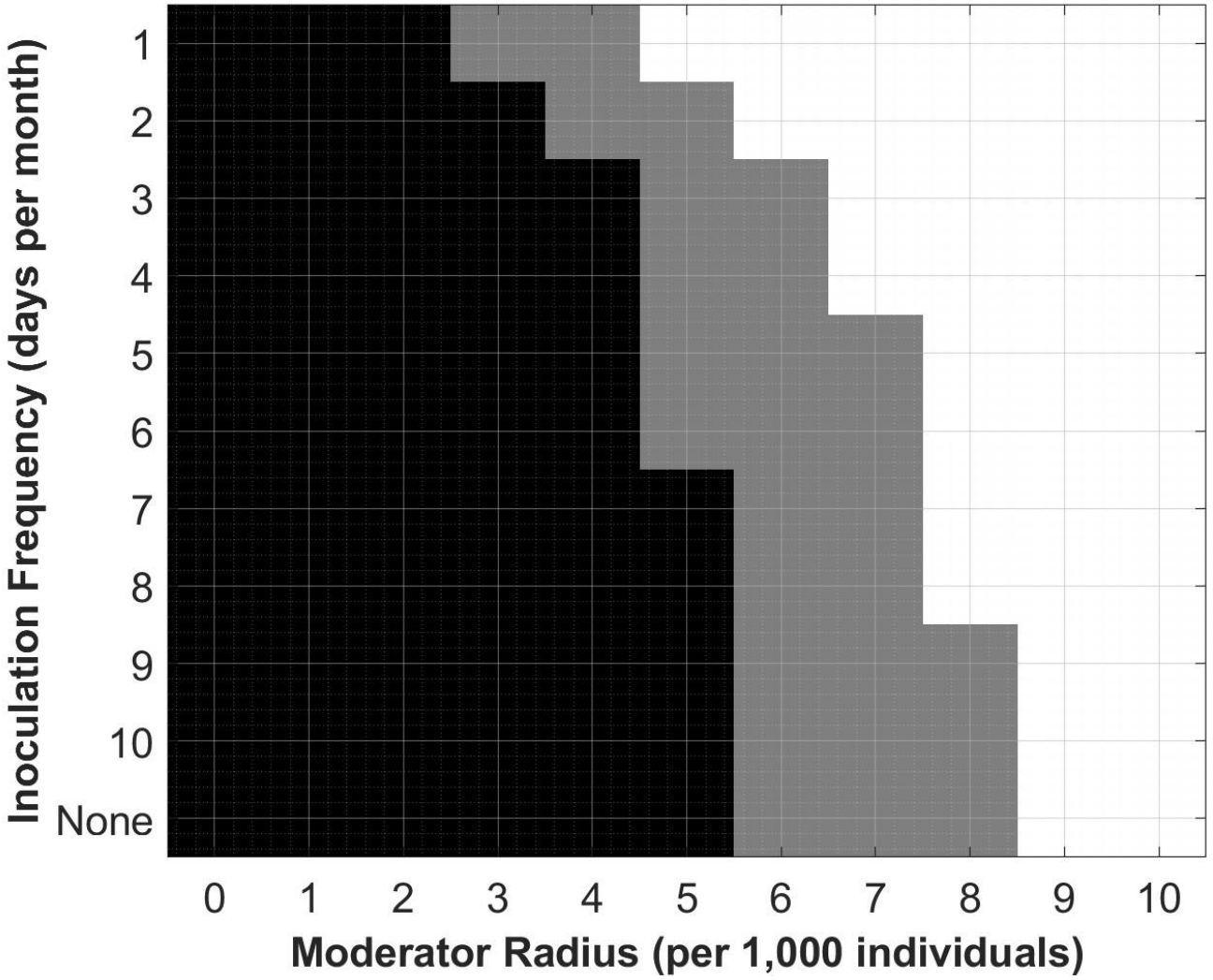
- From 1 training event every 10 days to 1 training event daily [11 possible levels of inoculation]

Hence, policy space is made of $\mathbf{11 \times 11 = 121}$ **policy baskets**

- For each mix, model is simulated for $\mathbf{T = 365}$ **periods** and replicated **across** $\mathbf{M = 100}$ **runs.**
- The procedure generates a pool of $11 \times 11 \times 100 \times 365 = 4{,}416{,}500$ artificial data points.

Performance measure: **end-simulation number of radicalized users**

Result 1: Performance Matrix & Substitution

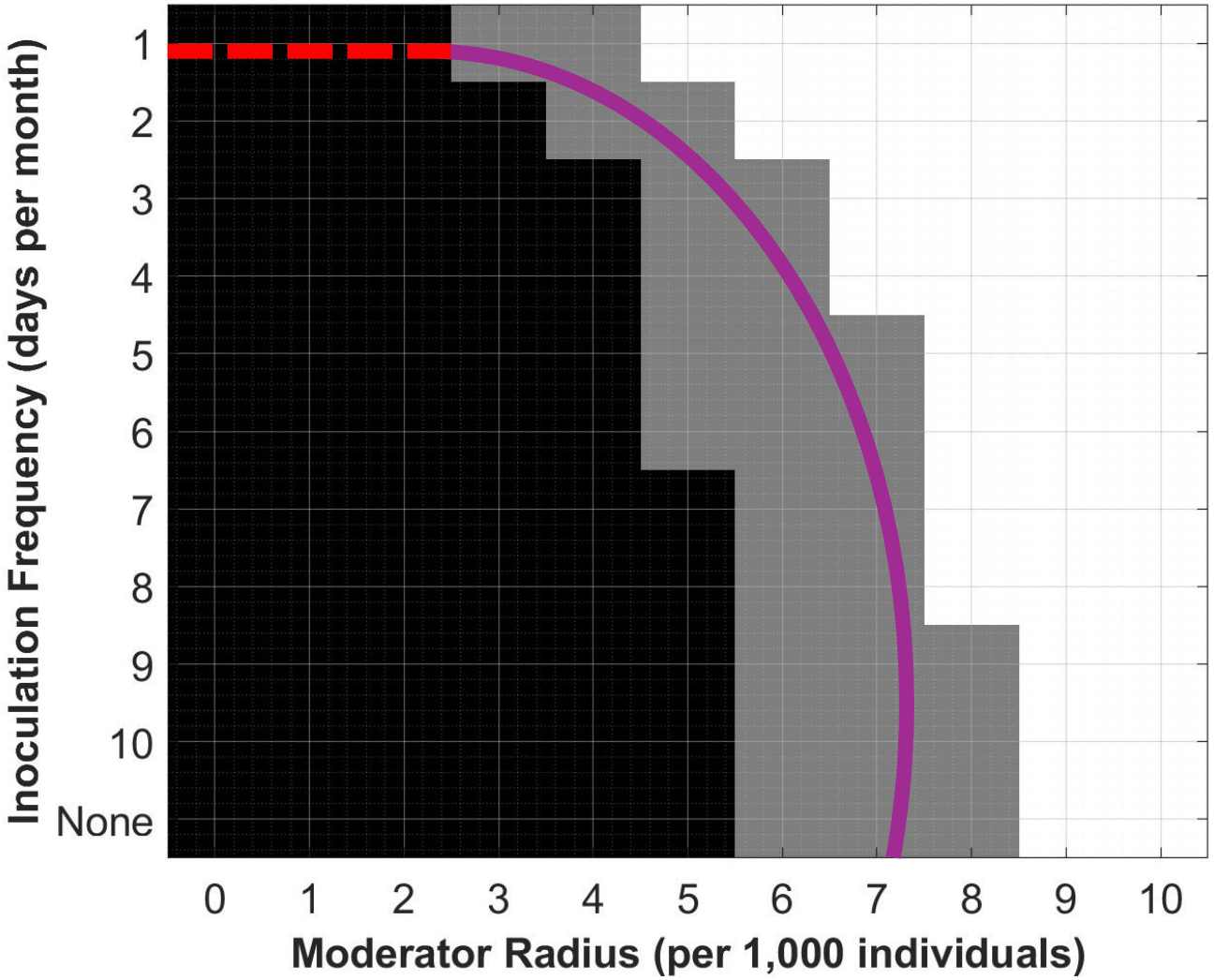


Performance measure:
End fraction of radicalized users

Policy baskets can be grouped in:

Low Performance (black area):
Above **5%**
Medium Performance (grey area):
Between **3%** and **5%**
High Performance (white area):
Below **3%**

Result 1: Performance Matrix & Substitution

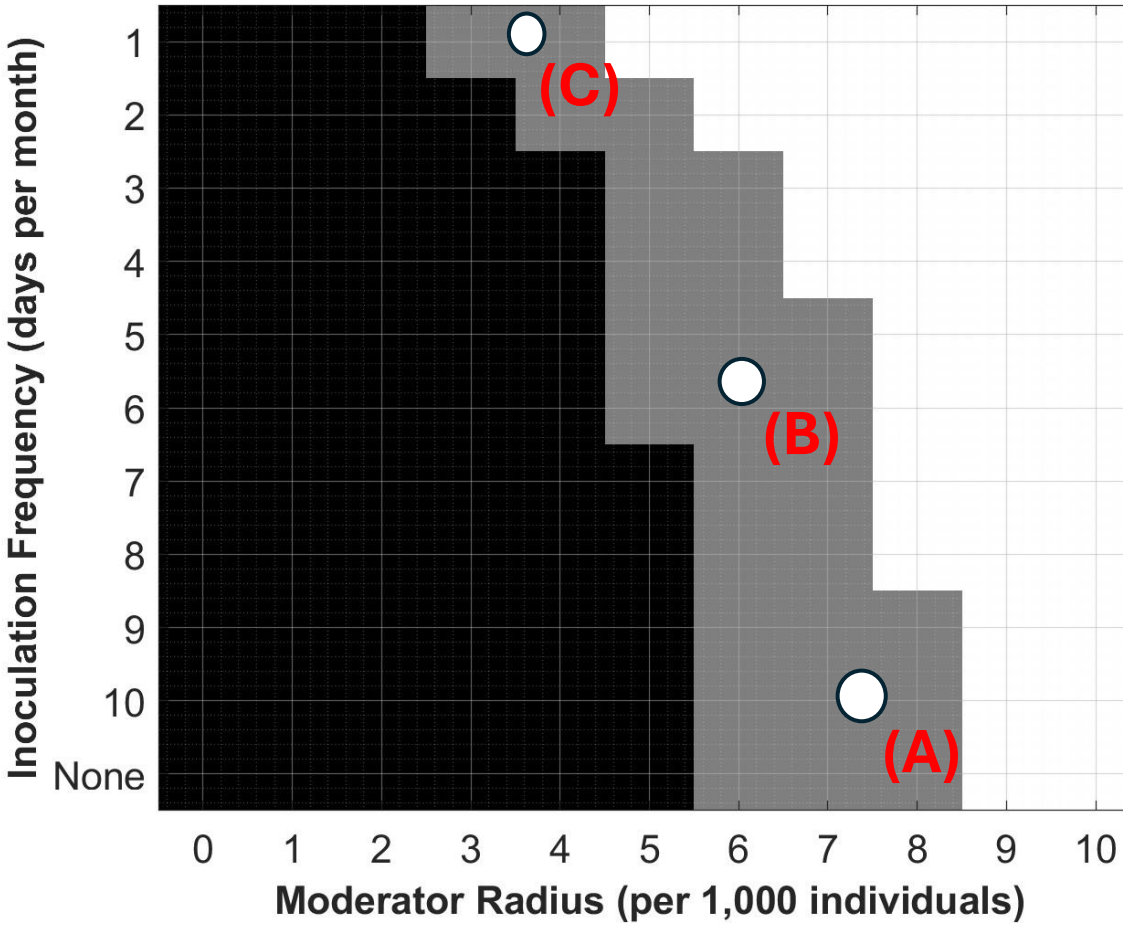


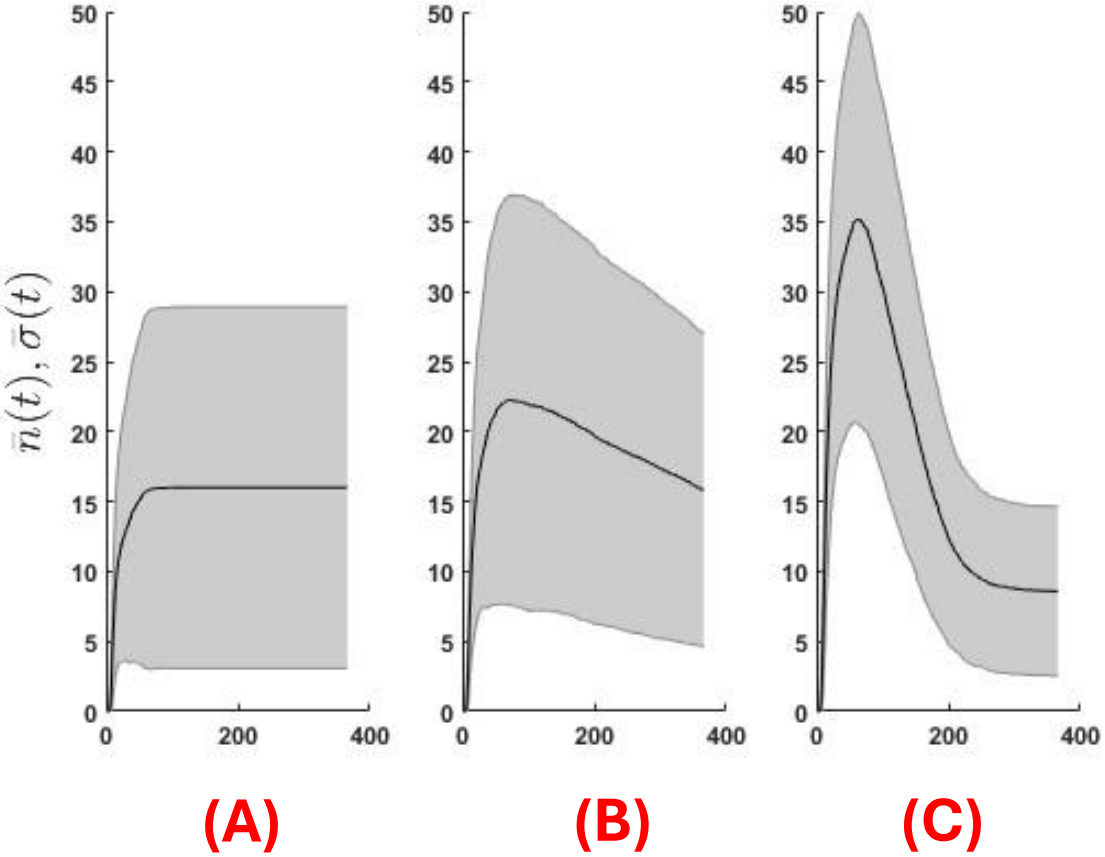
To some extent, policy-makers can substitute repression with inoculation

However, to attain medium and high performance, **some** level of moderator activity is **always** required!

This means that policies are only **partial substitutes**

Result 2. Everything equal, inoculation **reduces volatility**.





Result 4. Cost-benefit analysis.

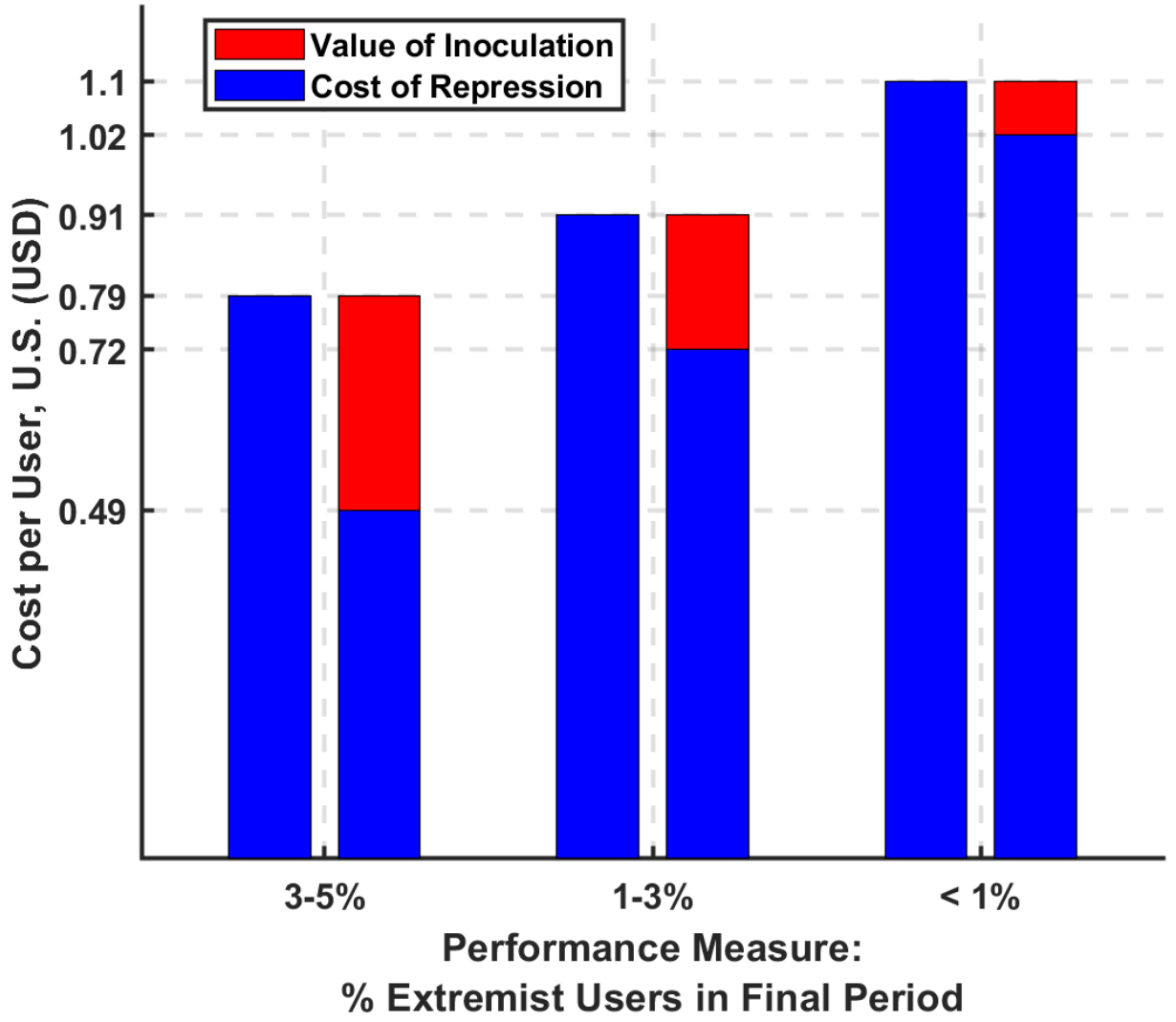
Example with real-life U.S. data (Perrigo, 2022)

Facebook moderator's productivity.

- Average Handling Time: 50s/item
- A single Facebook contractor requires 10 days to investigate 1% of the user population

Cost structure:

- E.g. a single Facebook contractor monitoring 1% of the population costs **$459.90**



Take-Home Messages

1. We develop a **toolbox** which is **general in purpose** as it allows to test for **multiple policies** in isolation and in combination for multiple (and varying!) institutional and demographic settings
2. We find that **to some extent**, inoculation decreases the need (**and the cost**) of moderation.
3. **Inoculation works by making the societal orbit more predictable and society more resilient.**
4. The model can be used to run a first-pass cost-benefit analysis for policies that are hard to assess in real data.

Risk state	**Immune**	**Low-risk**	**Susceptible**	**High Risk**	**Extreme**
Radicalisation State	***Reject***		*Uncertain*	***Accepts***	
Radicalisation Score (B)	$B < 0.05$	$0.05 < B \leq 0.33$	$0.33 < B \leq 0.66$	$0.66 < B \leq 0.95$	$B > 0.95$