



Behavioral
Data Science

From individual actions to
collective phenomena

Data Science Institute



Dr Marian-Andrei Rizoiu | Behavioral Data Science Lead
Marian-Andrei.Rizoiu@uts.edu.au
<https://www.behavioral-ds.science>



UTS CR

Who are our **digital** opinion leaders?



Jenna Abrams

@Jenn_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennabrams@gmail.com (Yes, there are 3 Ns, this is important)



jennabrams.com

Joined October 2014

Born on October 02



Tennessee GOP

@TEN_GOP

I love God, I Love my Country

Tennessee, USA

Joined November 2015

Common traits:

- Highly influential;
- Highly followed and retweeted;
- Opinion leaders;
- ...



General Flynn

@GenFlynn

Believe in #AmericanException
it's for real. NYT bestselling au
@fieldoffight (Read it) I stand i
American principles & values.

GLOBAL

mikeflynndefensefund.org

S-a alăturat în Ianuarie 2014

Sebastian Gorka DrG

@SebGorka

@FOXNews Nat. Sy. Strategist
Bestseller: DEFEATING JIHAD amzn.to/2zTuXyl
Fmr Strategist to Pres.
Trump Order my NEW book: amzn.to/2JkuSGJ

Washington, DC

S-a alăturat în martie 2014

Who are our **digital** opinion leaders?



Jenna Abrams

@Jenn_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennabrams@gmail.com (Yes, there are 3 Ns, this is important)

📍 USA

🔗 jennabrams.com

📅 Joined October 2014

📍 Born on October 02



Tennessee GOP

@TEN_GOP

I love God, I Love my Country

📍 Tennessee, USA

📅 Joined November 2015

Common traits:

- Highly influential;
- Highly followed and retweeted;
- Opinion leaders;
- ...

Russian-controlled trolls
operated by the Internet Research Agency in St. Petersburg

Three rules of bird flocking behaviour



Complex bird flock dynamics...

Three rules of bird flocking behaviour



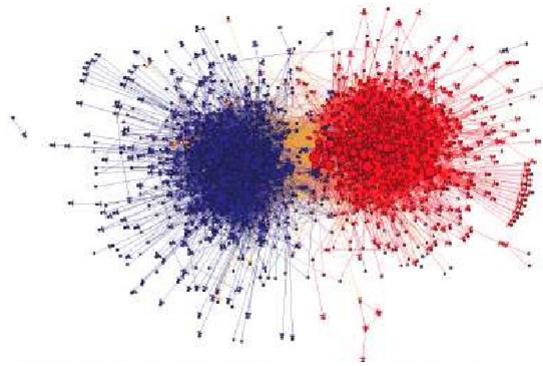
Complex bird flock dynamics...



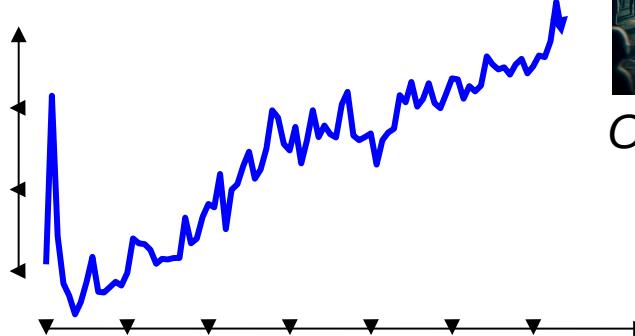
... can be explained by three individual level rules:

1. Separation
2. Alignment
3. Cohesion

Our individual online actions shape online phenomena



Opinion polarisation



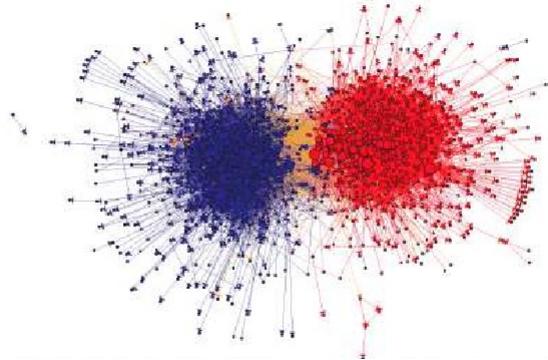
Popularity, virality, attention



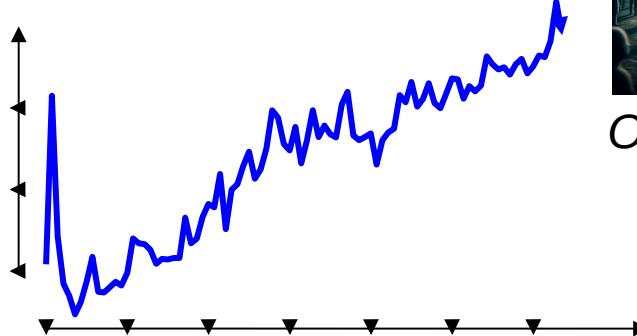
*Opinion manipulation
Digital propaganda*

Complex societal phenomena ...

Our individual online actions shape online phenomena



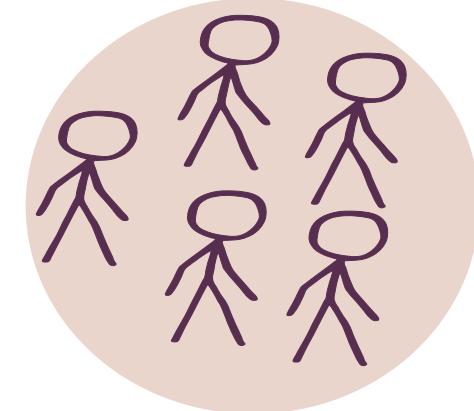
Opinion polarisation



Popularity, virality, attention



*Opinion manipulation
Digital propaganda*



... are shaped by the individual-level decisions that people make:

- What we reshare
- How we find our peers
- How we consume information
- What shapes our values

What if...

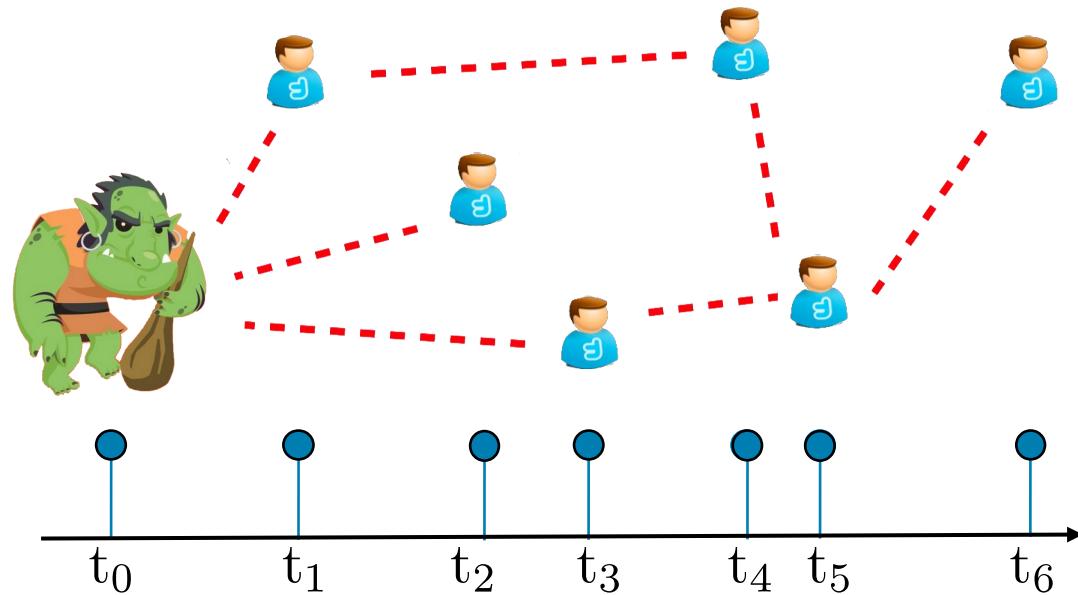
... we could understand and model individual level actions?

... we could predict the societal level effects?

... we can map in realtime information flow at scale?

... we could quantify the effectiveness of our (counter)messaging campaigns before deployment?

Today: Detect Information operation agents



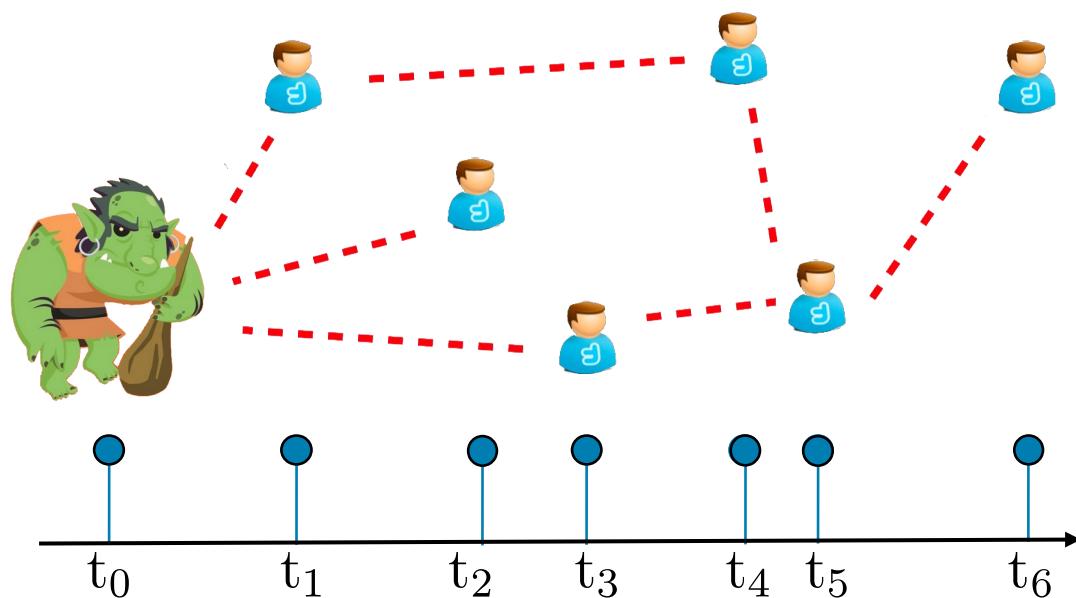
We can use the reactions of social systems ...



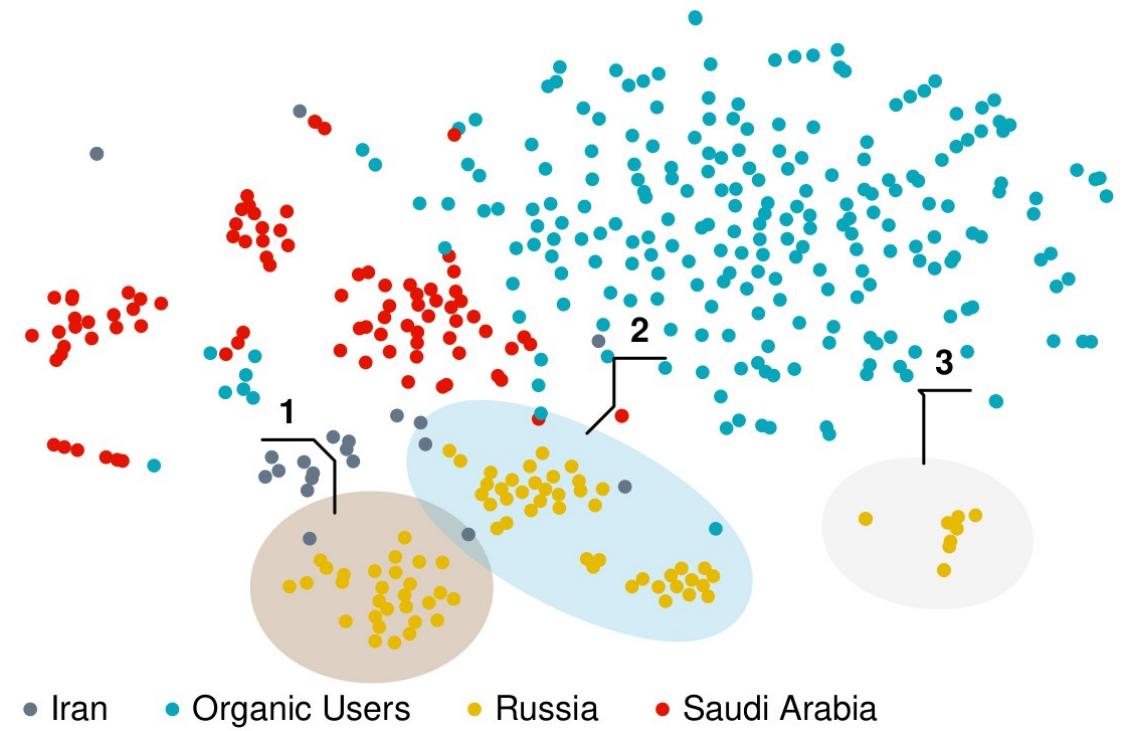
The technical detail:

Users as social sensors; quantify users reaction to agents; Hawkes modelling; Transformer technology

Today: Detect Information operation agents



We can use the reactions of social systems ...



... to detect Information Operations agents and even individual troll farms.



The technical detail:

Users as social sensors; quantify users reaction to agents; Hawkes modelling; Transformer technology

Today: Detect inauthentic coordinated campaigns



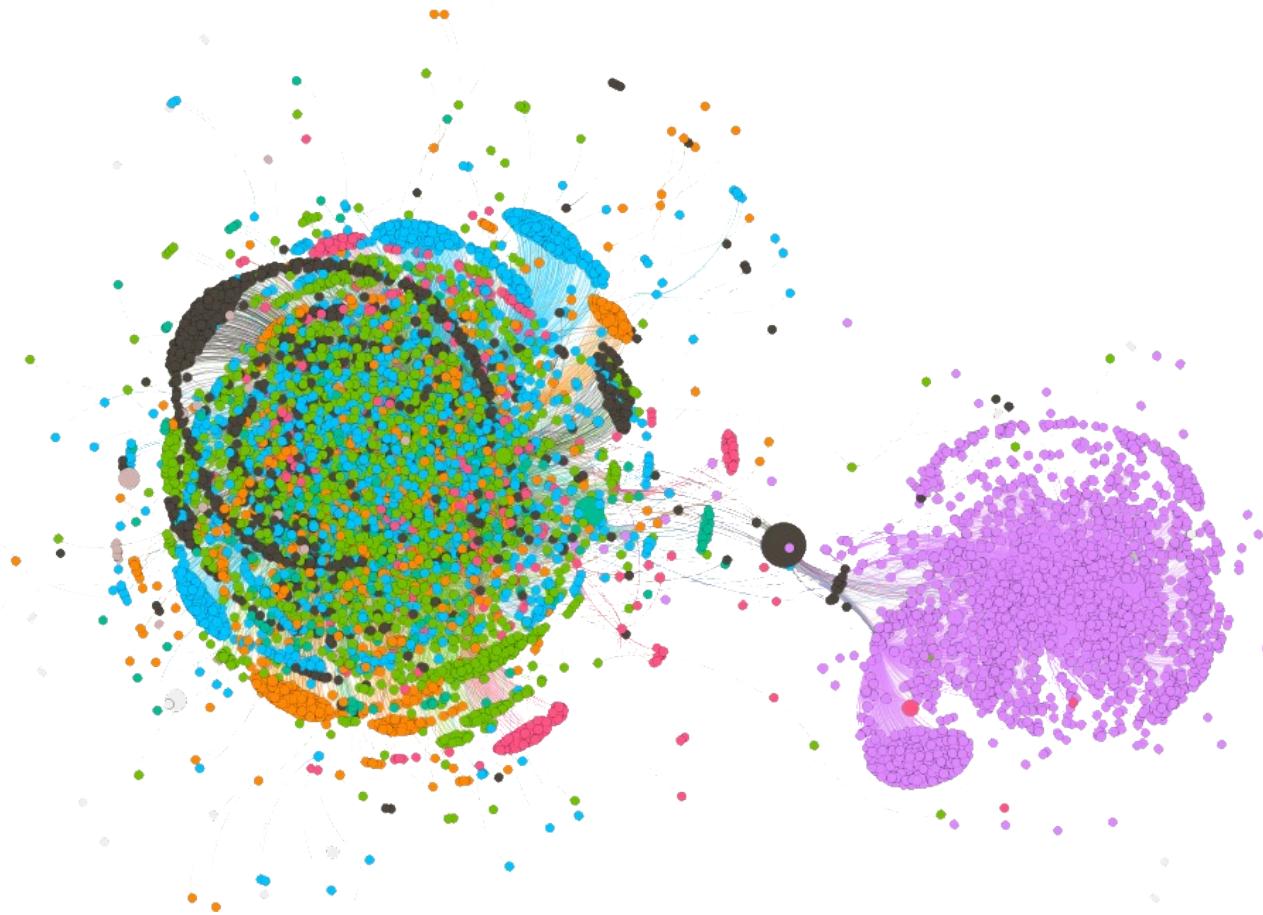
The structure of the information spread networks ...



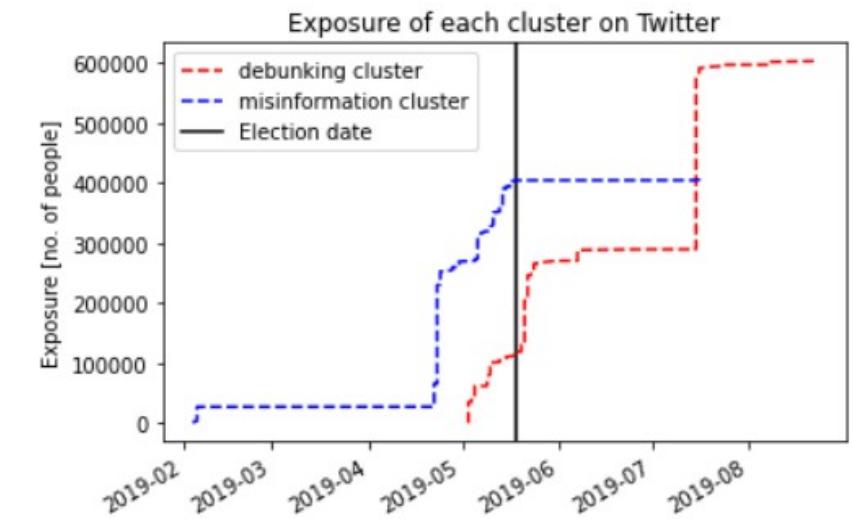
The technical detail:

Map information networks from social media; content, interactions, structure and diffusions analyse; social network analysis

Today: Detect inauthentic coordinated campaigns



The structure of the information spread networks ...



... can predict coordinated interventions



The technical detail:

Map information networks from social media; content, interactions, structure and diffusions analyse; social network analysis

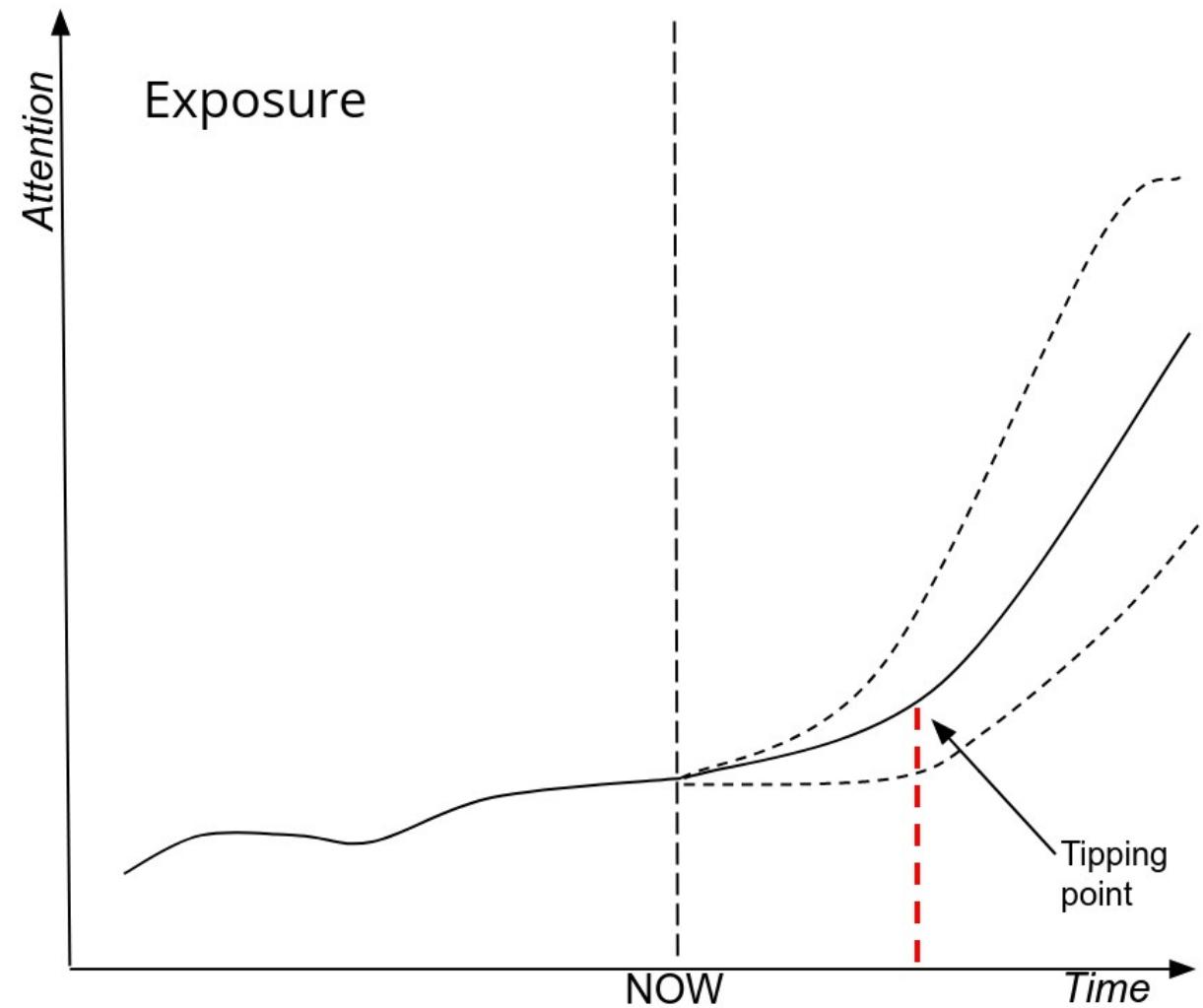
Soon: Predict the effectiveness of IO campaigns

Given an early observed spread, predict the future attention a narrative will gain

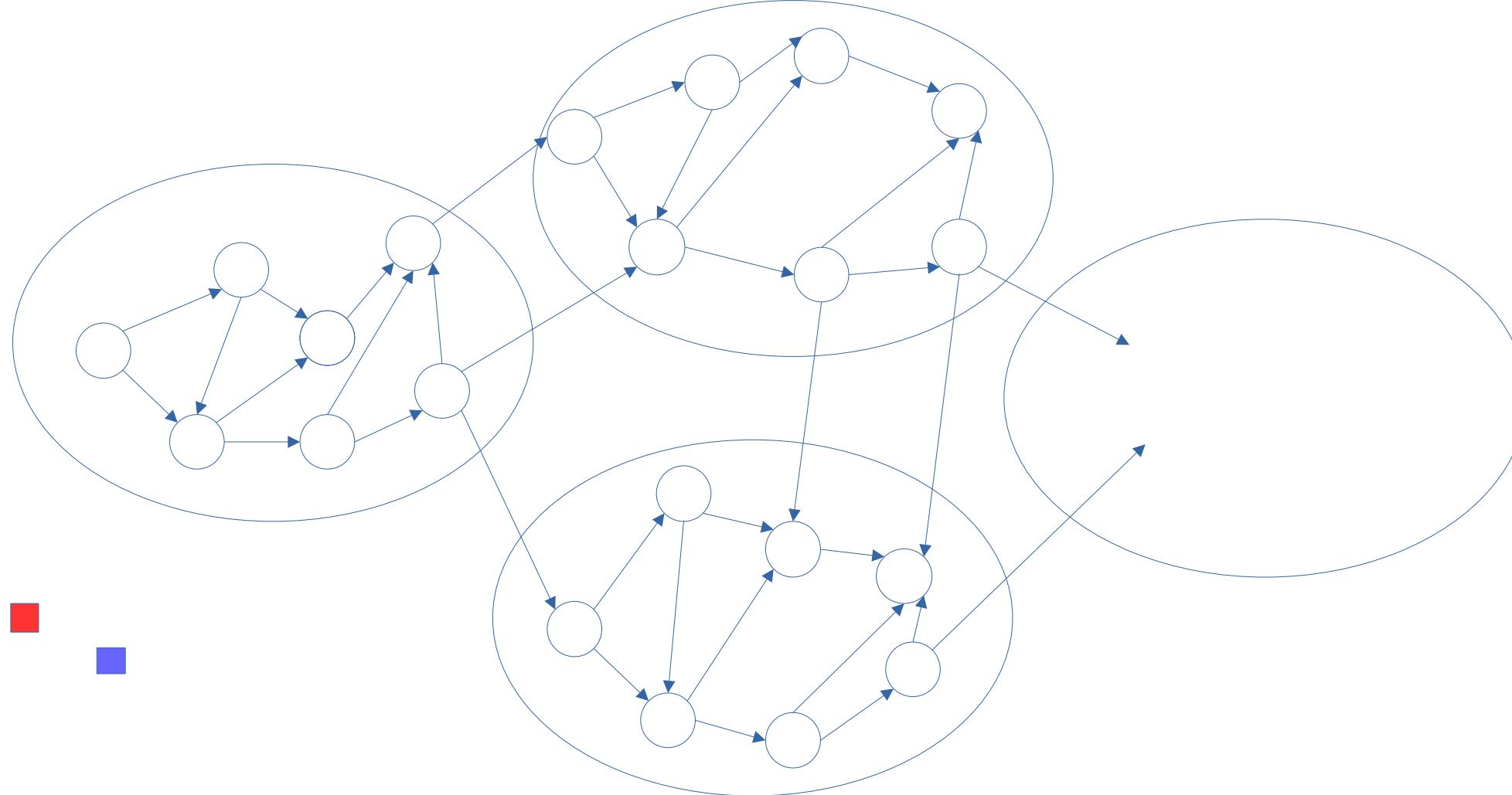
tipping points – the points at which the reach of IO increases exponentially

When to act? (**hint:** before tipping point)

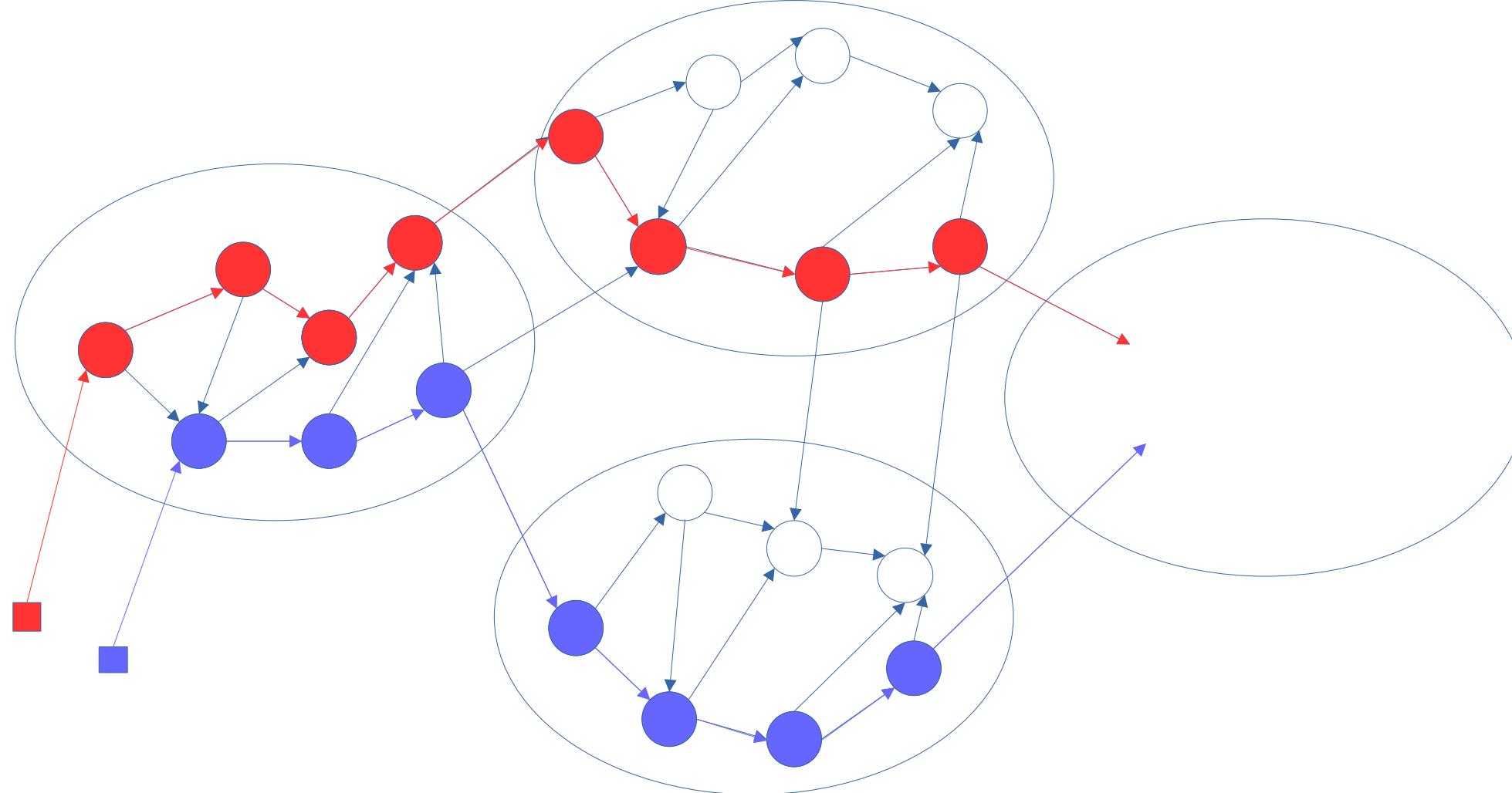
What to counter? (**hint:** potentially popular IO)



Soon: Predict the “path” of an information spread



Soon: Predict the “path” of an information spread



Research plan in Information and Influence



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Research plan in Information and Influence

Thank you!



Response
level



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

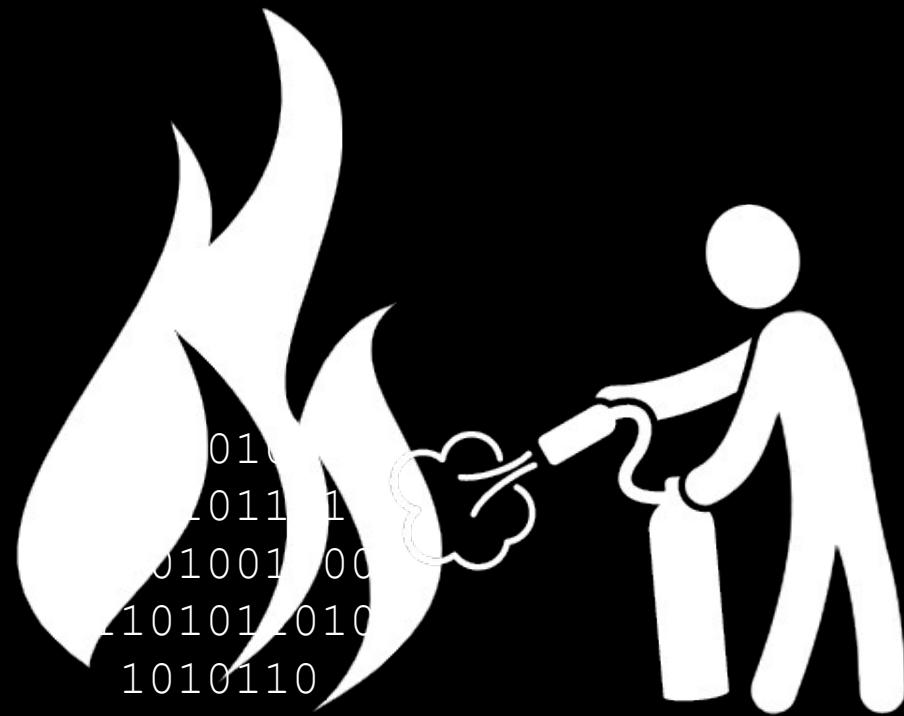
Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

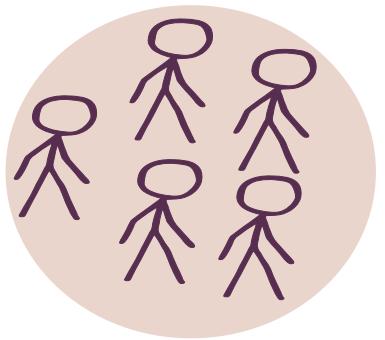
Identify key message inoculation points in social networks based on how information flows and gains velocity



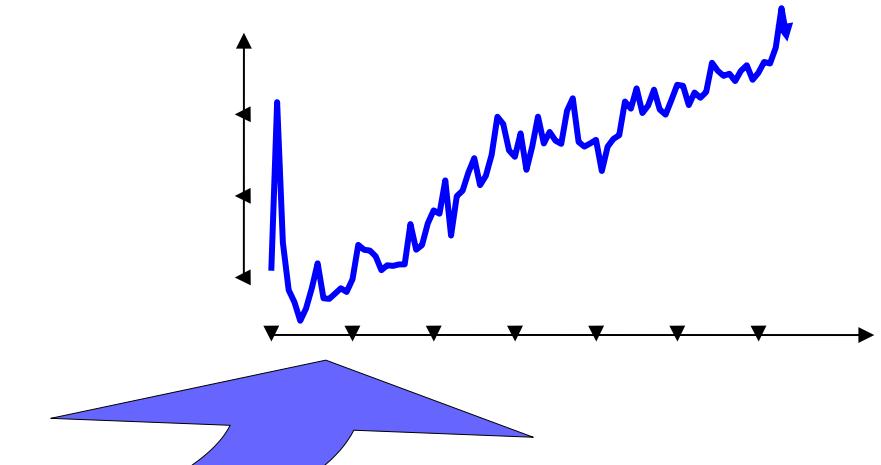
Other capabilities & expertise

The Behavioral Data Science

1.

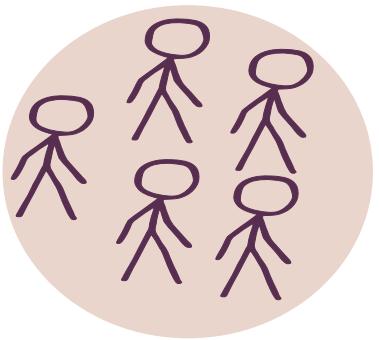


information diffusion
epidemics spreading
behavioral modeling

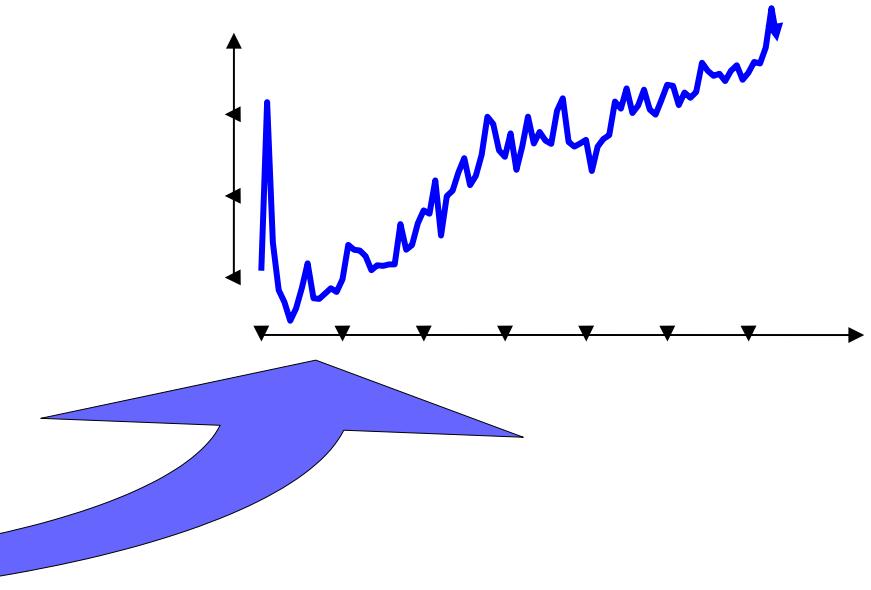


The Behavioral Data Science

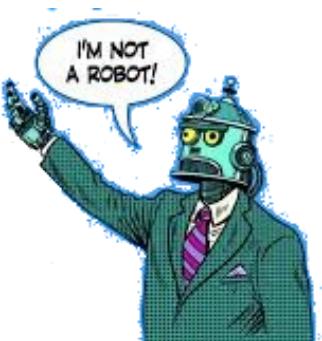
1.



information diffusion
epidemics spreading
behavioral modeling



2.

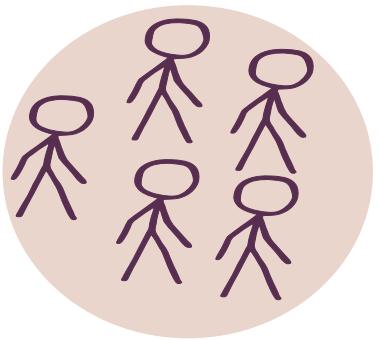


[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

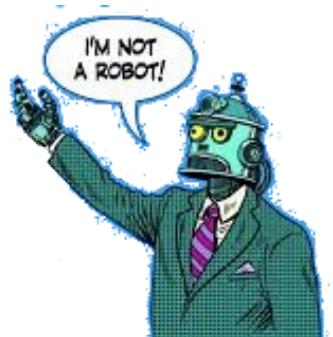
The Behavioral Data Science

1.



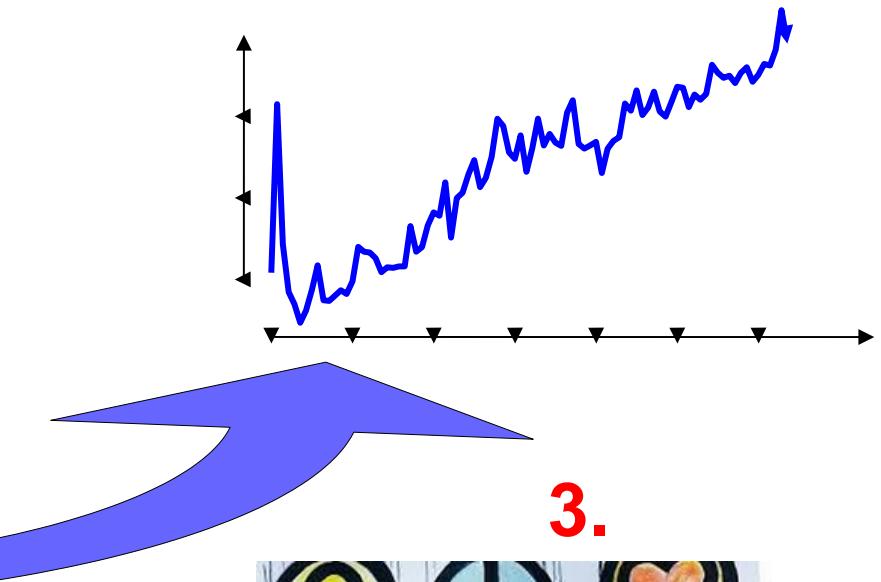
information diffusion
epidemics spreading
behavioral modeling

2.



[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]



3.



More challenges: Red Queen effect and generative models

Content- and user-based detection tools:

- language nuances, language drift, foreign language-dependent
- sensitive to adversarial training attacks – use the detector to train the attacker.

ChatGPT-like introduced problems:

- **Increase scalability of troll presence** – # accounts manned by a single operator increase multifold
- Increased awareness of **social context knowledge**, adapted to local language patterns
- **Adaptability for detection** – generative models can spit out infinite variations of the same message.



"Now, here, you see, it takes all the running you can do just to keep in the same place. If you want to get somewhere else, you must run at least twice as fast!"

- Red Queen from *Through the Looking Glass*

Solution: not what they say, but what they stir

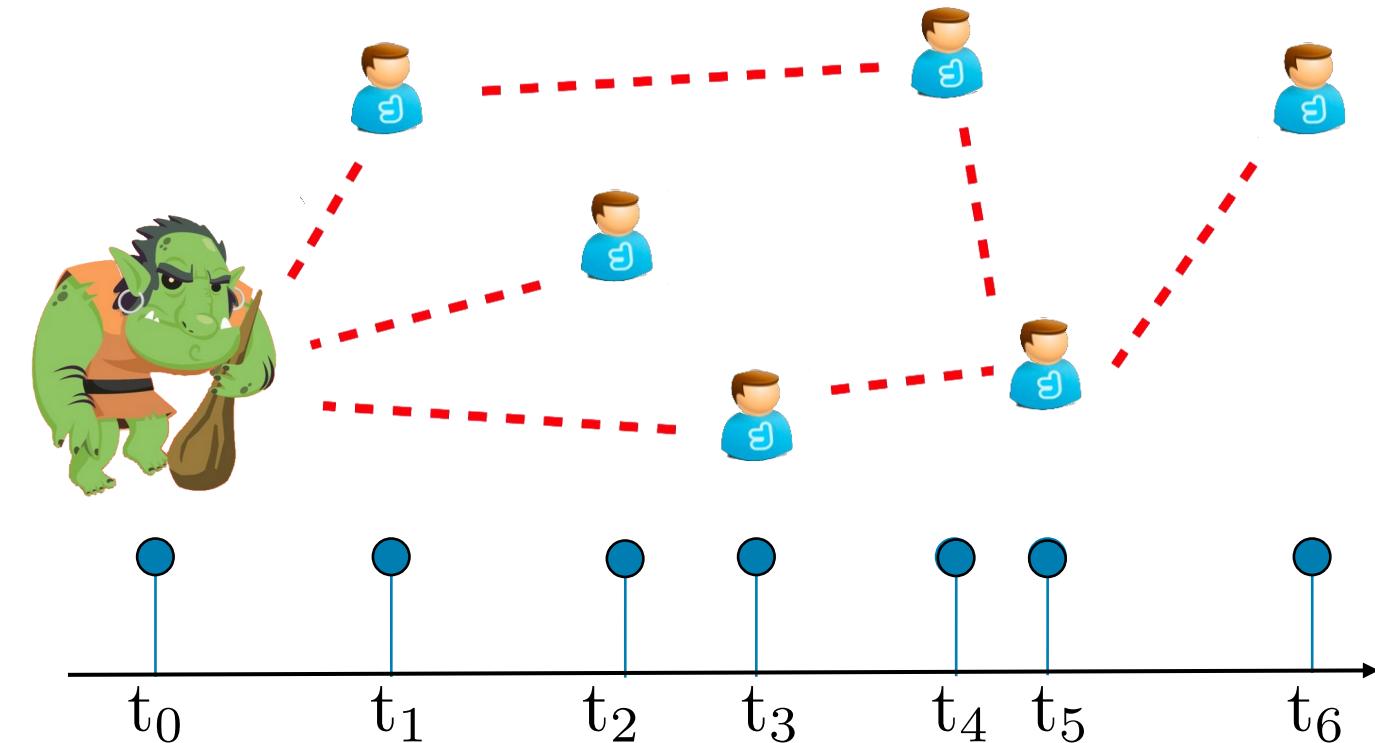
IO are designed to elicit particular reactions from the target audience

- use these reactions to detect
- cannot be faked or hidden, as they are the purpose of the campaign

Distinguish users and content types
based on the reaction of online
social systems (**no content**)

Build early detection systems based
on information spread patterns within
the user population

Language invariant – the model
doesn't need to “speak” the language



Our founders (and friends) in the mis-, dis-, IO and IW spaces



Australian Government
Department of Defence
Defence Science and
Technology Group

Modelling in the Gray Zone: Complex
Warfare “Real-time detection of
disinformation campaigns”



Australian Government
Department of Defence
Defence Science and
Technology Group

Information Warfare STaR
Shot “NARRATE: Developing
Situational Awareness”



Information integrity initiative:
fighting misinformation in Australia



Effectiveness of Information
Operations in the Solomon
Islands and Melanesia



WIKIMEDIA
FOUNDATION



WIKIPEDIA
The Free Encyclopedia

Improving Wikimedia resilience against
the risks of content-generating AI systems



Hate Speech propagation
on Social Media

Behavioral DS action plan in Influence Operations space



**Response
level**



Objective



Approach

Monitor

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

What are practical approaches that allow us to both proactively and reactively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Action plan



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Monitor: Monitoring discussion spaces

Information Dashboard

Segments

- Misinformation
- The Voice
- All Content Streams
- [See all Segments >](#)
- [See all Content Streams >](#)

Topics

- #The Voice
- #Test
- #Kherson**
- [See all >](#)

Analytics

Search... Add Topic

Filter/Sort Saved Views

#Kherson

Description 1

 **Kyiv Independent** https://www.instagram.com/kyivindependent_official/ 10 Sep / 16:12

The headquarters of the political party "United Russia" located in the occupied city of Polohy, Zaporizhzhia Oblast, were destroyed on Sept. 8, Melitopol Mayor Ivan Fedorov reported via Telegram. Fedorov said local...

 **Institute for the Study of War** <https://www.facebook.com/265135661809> 10 Sep / 08:23

NEW: Ukrainian forces confirmed advances in the Donetsk-Zaporizhzhia Oblast border area and in western Zaporizhzhia Oblast and claimed advances south of Bakhmut on September 9. Russia's war in Ukraine is increasingly constrainin...

 **Institute for the Study of War** <https://www.facebook.com/265135661809> 10 Sep / 08:23

NEW: Ukrainian forces confirmed advances in the Donetsk-Zaporizhzhia Oblast border area and in western Zaporizhzhia Oblast and claimed advances south of Bakhmut on September 9. Russia's war in Ukraine is increasingly constrainin...

 **EEAS** <https://www.instagram.com/eudiplomacy/> 09 Sep / 21:16

A reminder to Russia: Staging sham 'elections' will never justify the invasion or legitimise the illegal military control and temporary occupation of parts of Ukrainian territories. Kherson, Zaporizhzhia, Donetsk, Luhansk, and Crimea...

 **Institute for the Study of War** <https://www.instagram.com/thestudyofwar/> 09 Sep / 21:10

Ukrainian forces continued offensive operations near Bakhmut and in western Zaporizhzhia Oblast on September 7 and made further gains on both

 **Institute for the Study of War** <https://www.instagram.com/thestudyofwar/> 09 Sep / 00:10

Midday #Ukraine Map Update: Ukraine has made further gains in western #Zaporizhzhia. Occupied frontlines moved September 7 above that Ukrainian

<< < 1 2 3 4 5 6 7 > >>

23896 total posts. Analyse posts

Summary

Total 23896

Source	Posts	Value
Instagram	37	0.2%
Facebook	58	0.2%
Twitter	23800	99.6%
Telegram	1	0.0%

Keywords

Мелітополь, Мелітополь, Melitopol, Ніколаїв, Mykolaiv, Вознесенськ, Вознесенськ, Voznesensk, Запорож'я, Zapоріжжя, Enerгодар, Enerhodar, Khersonoffensive, Khersoncounteroffensive, Херсон, Kherson



Action plan



Response level



Objective



Approach

Monitor

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

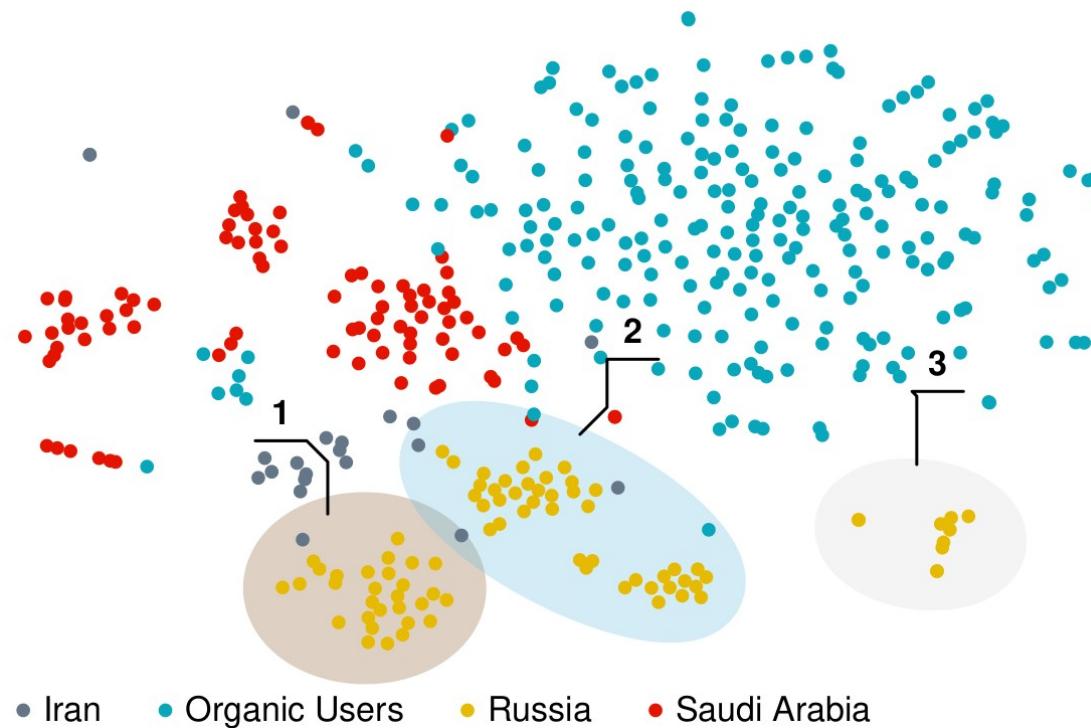
What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Detect: identify agent types and coordinated behavior



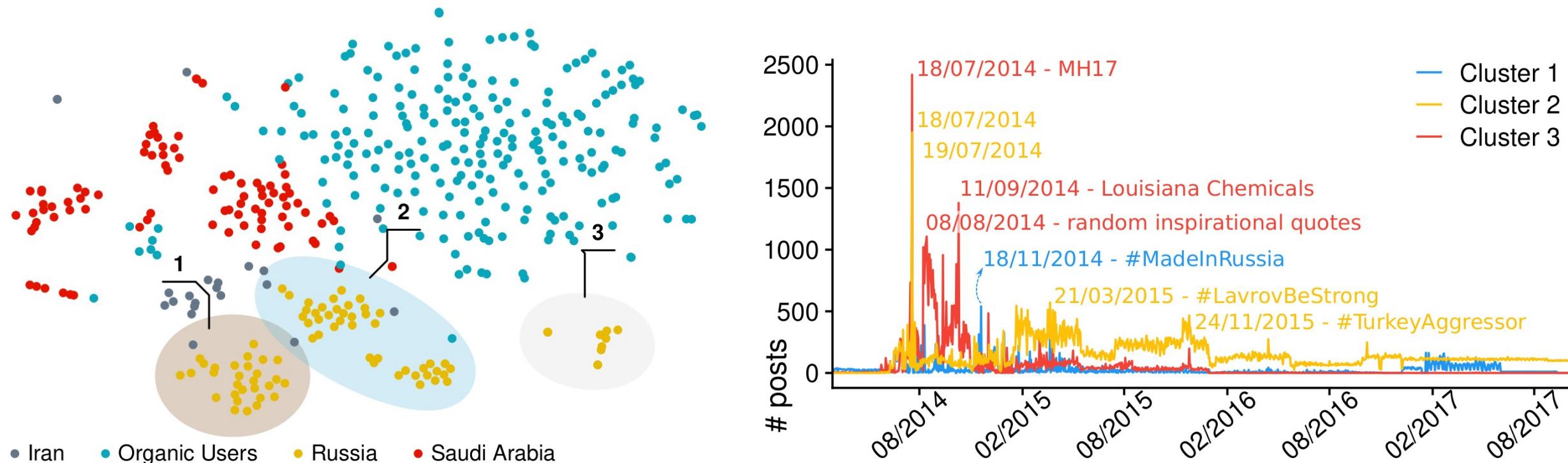
IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.



The technical detail:

Interval-censored Transformer Hawkes; Twitter Moderation Research Consortium dataset; partial data loss

Detect: identify agent types and coordinated behavior



IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.

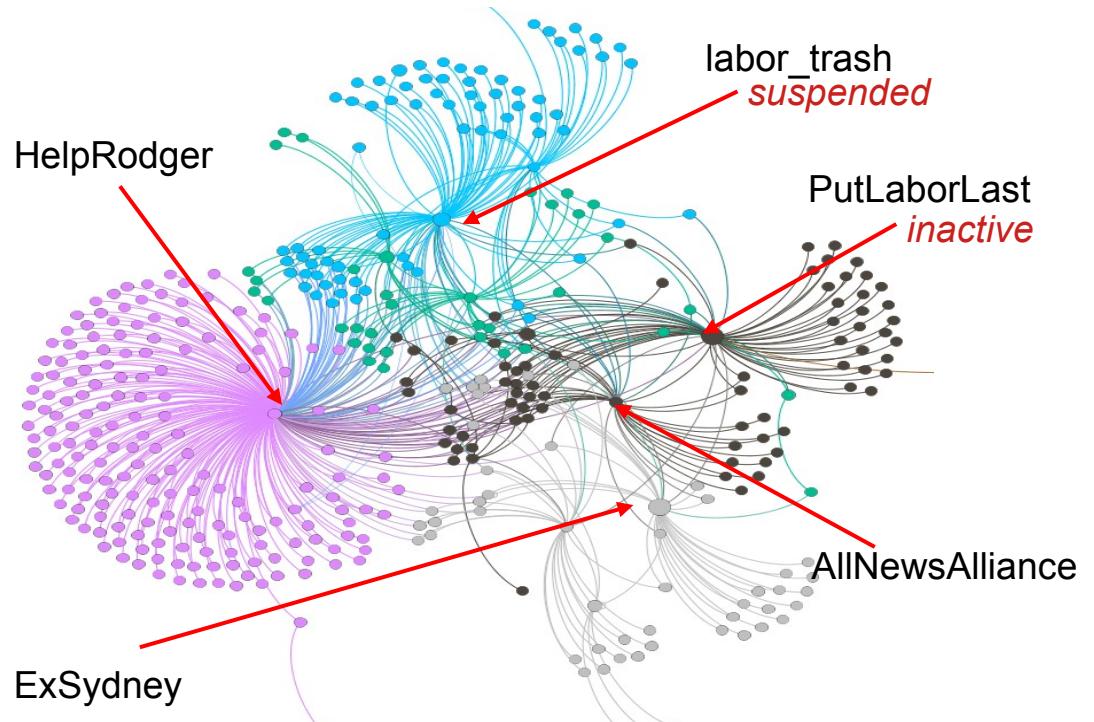
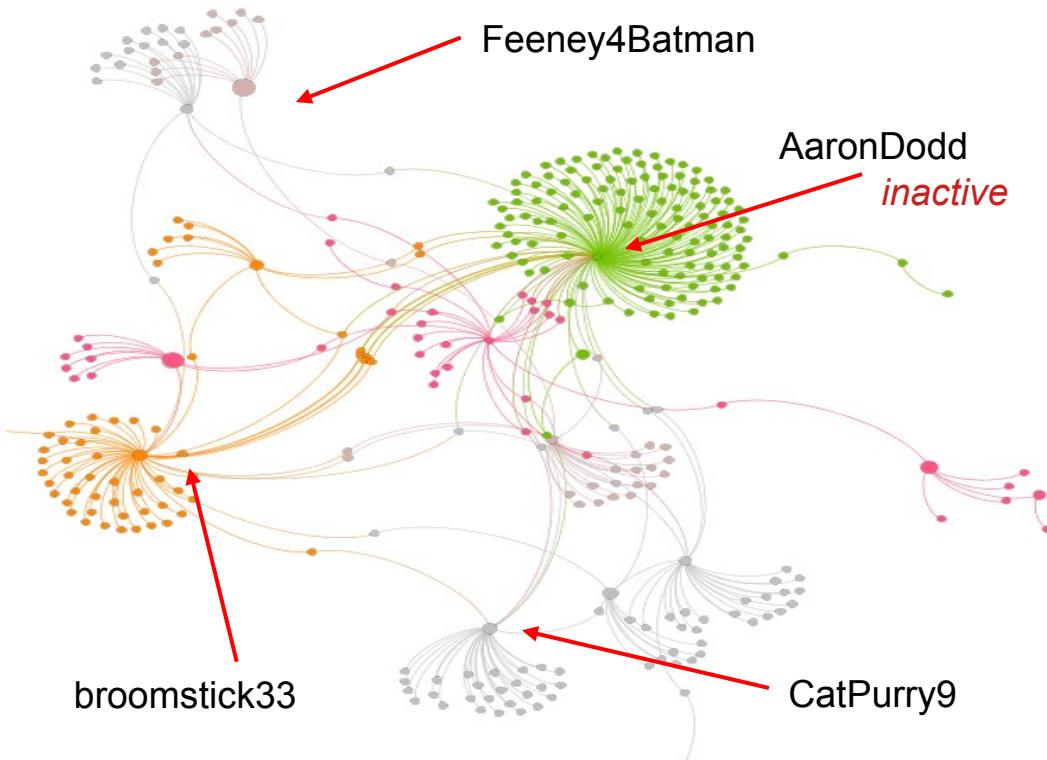
Qualitative investigations uncovers strategies of Russian trolls farms:
C1: Russian news with patriotic framing;
C2: Regional and conservative news;
C3: tweet in English, *#music*, *#usa*, relationship advice



The technical detail:

Interval-censored Transformer Hawkes; Twitter Moderation Research Consortium dataset; partial data loss

Detect: two clusters emerge



Observations:

- Two clusters – one misinformation and one debunking the misinformation
- Misinformation cluster (right) is strongly connected compared to debunking cluster (left), cross resharing



Detect: Detect Russian IO operatives and their narratives in Ukraine

Collect data on Telegram and Twitter, in three languages (Russian, Ukrainian and English)

Kherson counteroffensive (29/08/2022- 11/11/2022):

#Kherson / #Херсон
#Khersoncounteroffensive
#Khersonoffensive
#Enerhodar / #Енергодар
#Zaporizhia / #Запоріжжя / #Запорожье
#Voznesensk / #Вознесенськ / #Вознесенск
#Mykolaiv / #Миколаїв / #Николаев
#Melitopol / #Мелітополь / #Мелитополь

Kharkiv counteroffensive (06/09/2022-02/10/2022):

#Kharkiv / #Харків / #Харьков
#Izium / #Ізюм / #Изюм
#Balakliia / #Балаклія / #Балаклия
#Kupiansk / #Купянськ / #Купянск
#Sviatohirsk / #Святогірськ / #Святогорск
#Svatove / #Сватове
#Kremenna / #Кремінна / #Кременная
#Luhansk / #Луганськ / #Луганск
#Shevchenkove / #Шевченкове
#Verbivka / #Вербівка
#Yakovenkove / #Яковенкове



Detect: Detect Russian IO operatives and their narratives in Ukraine

Information Dashboard

Segments

- Misinformation
- The Voice
- All Content Streams

See all Segments >

See all Content Streams >

Topics

- #The Voice
- #Test
- #Kherson

See all >

Analytics

- Kherson
- Report: 2023/09/04 11:44
- Report: 2023/07/28 12:20

See all >

Kherson

Kherson

21/06/2023 - 23/06/2023 Twitter

View more details

Hover to view user or create a cluster

Create Cluster

View By: Unclustered War Narrators The Vladolf clan The Falconers

Edit Clusters

Post Frequency

Number of Posts

Post Frequency

1AM Jun 22nd 2AM Jun 22nd 3AM Jun 22nd 4AM Jun 22nd 5AM Jun 22nd 6AM Jun 22nd 7AM Jun 22nd 8AM Jun 22nd 9AM Jun 22nd 10AM Jun 22nd 11AM Jun 22nd 12PM Jun 22nd 1PM Jun 22nd 2PM Jun 22nd 3PM Jun 22nd 4PM Jun 22nd 5PM Jun 22nd 6PM Jun 22nd 7PM Jun 22nd 8PM Jun 22nd 9PM Jun 22nd 10PM Jun 22nd 11PM Jun 22nd 12AM Jun 23rd 1AM Jun 23rd

Minimum Connections:

Labels

Narrative Summary - Enerhodar

Export Summary

Export Posts

Detect: Detect Russian IO operatives and their narratives in Ukraine

Information Dashboard

Segments

- Misinformation
- The Voice
- All Content Streams
- See all Segments >
- See all Content Streams >

Topics

- #The Voice
- #Test
- #Kherson
- See all >

Analytics

- Kherson
- Report: 2023/09/04 11:44
- Report: 2023/07/28 12:20

See all >

Kherson

View By: Unclassified

Number of Posts

1AM Jun 22nd 2AM Jun 22nd 3AM Jun 22nd 4AM Jun 22nd 5AM Jun 22nd

Narrative Summary

+ - Hover to view bio or click to add content stream to a cluster.

Selection Tools

Kherson

21/06/2023 - 23/06/2023 Twitter

23800 posts retrieved. Keywords include Мелітополь, Мелітополь, Melitopol, Nikolaev, Mykolaiv, Voznesensk, Вознесенськ, Voznesensk, Zapорожье, Запоріжжя, Zaporizhia, Enerгодар, Enerhodar, Khersonoffensive.

[View less details](#)

Cluster Name:

Users Selected

MR. M #Ghost Of No...

Reset Create Cluster

Kherson

21/06/2023 - 23/06/2023 Twitter

View more details

Hover to view user or create a cluster

Create Cluster

Edit Clusters

Labels

Export Summary

Export Posts

Action plan



**Response
level**



Objective



Approach

Monitor

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Predict: Effectiveness of state-sponsored narratives

Purpose: Predicting the Effectiveness of State-sponsored Influence Operations: a Case Study of the Solomon Islands and Melanesia

Questions:

RQ1: Do Chinese IO operations in different regions of Melanesia consistently use particular narrative constructions and language? Do they seek to elicit specific reactions?

RQ2: How to effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

RQ3: What factors accelerate the reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harm?

RQ4: How to forecast the future attention received by an IO? How to distinguish those IO likely to be effective? Is there a tipping point before which countermeasures are most effective?

Predict: Melanesia IO – on-the-ground & protocol

Qualitative protocol:

- Compile a list of significant past events and study who pushed what
- Identify where the discussions take place
- Who are the actors (vectors) pushing narratives?
- What are the main themes and topics?

Two Research Assistants on the ground:

- **Places:** Solomon Islands, Vanuatu
- **Tasks:** Identify where are the discussions taken — on Facebook
- **Safety protocol:** they have limited knowledge of the aims of this project (i.e. China's IO).

The screenshot shows the CrowdTangle platform interface. At the top, there are navigation links: DASHBOARDS, LIVE DISPLAYS, INTELLIGENCE, and SEARCH. Below the navigation is a sidebar with sections for Notifications, Explore (which is selected), Lists, and a search bar. The Lists section includes options like Create List, MY FAVORITES (empty), PAGES (All Page Lists, Facebook Pages), GROUPS (All Group Lists, Facebook Groups), Saved Searches, and Saved Posts. The main content area is titled "All Page Lists" and shows a table of 12 Facebook pages. Each row includes a thumbnail, the page name, follower count, and three actions: Get History, Edit Details, and Remove. The pages listed are: 2023 Pacific Games Stadium Project (15,105 followers), ABC Pacific (413,893 followers), Audit the Audit Solomon Islands (784 followers), Chinese Embassy in Solomon Islands (6,061 followers), SIBC News (79,480 followers), Solomon Islands Parliamentary Opposition (17,515 followers), Solomon Star News (55,593 followers), and Solomon Times Online (19,813 followers).

Collecting Facebook data from identified vectors using CrowdTangle

Predict: Melanesia IO findings – Events and narratives

Five historical events in Solomons:

- Honiara Riots and burning of Chinatown
- Switch of public recognition from Taiwan to China
- Signing of the PRC-SI Security Pact
- Australia Offers to Fund Solomon Islands Elections
- U.S. embassy opening in Honiara

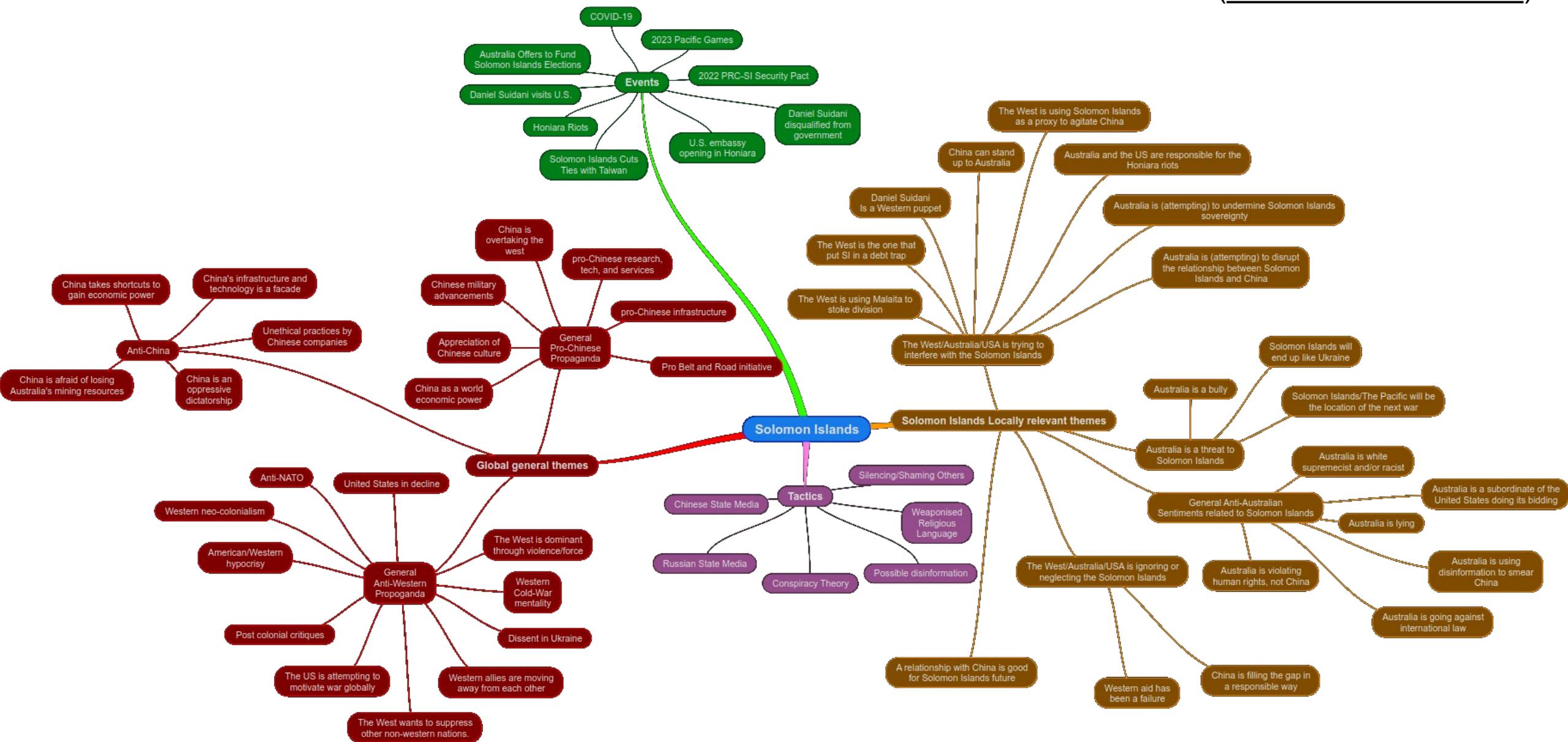
Predict: Melanesia IO findings – Events and narratives

(some) Identified narratives:

1. Solomon Islands should become economically self-reliant and vibrant post-RAMSI, through relations with China.
2. The West excludes China from the South Pacific to perpetuate their domination.
3. **Australia and the US are colonialist bullies that threaten Solomon Islands sovereignty.**
4. **The West perpetuate a ‘Cold War mentality’.**
5. Westernised activists and civil society groups undermine the Solomon Islands government.
6. Australia is trying to interfere in Solomon Islands democracy.
7. The West instigated the November 2021 Honiara riots.
8. **Australia and the West will invade the Solomons because of its security relations with China.**

Predict: Solomon Islands findings – Events and narratives

(zoomable version [here](#))



Predict: Melanesia IO findings – Network for fake accounts affiliated with China

- The implantation of narratives and false information is not in the stories, but in the comments
- Most of these comments are from a (small) set of key individuals
- We suspect that a number of these accounts are fake accounts – e.g., portraying to be honiarans when in fact they don't live in the Solomons
- The news articles do not really contain misinformation, only propaganda; some comments make fake claims
- We find that pro-chinese stories do not get a lot of engagement; local stories about Solomon's get engagement
- We find significant positivity about Australia in the people's comments
- Surveys show that ~80% of the elites in the Melanesia trust Australia – *where would you go for medical treatment or education?*

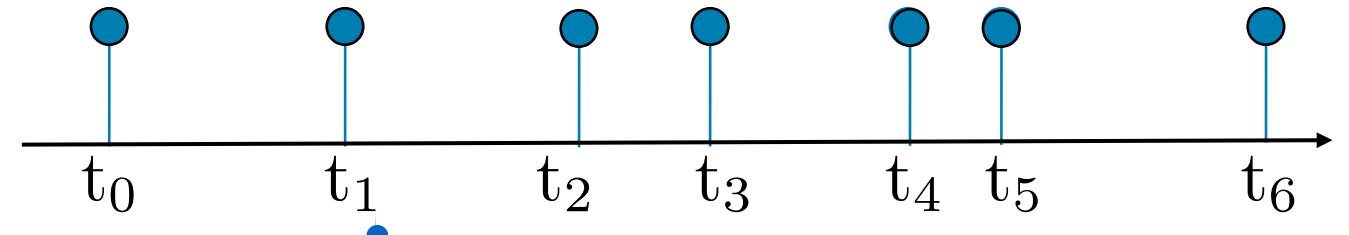
Predict: Melanesia IO findings – China exerts influence via trolling and policing on key issues

Hypothesis: China exerts influence via trolling and policing on key issues.

- when anti-China comments come in, the network of fake accounts barge in; usually using shaming
- usually the initial commenter fades back into the background
- Repetition is key – repeat the same messages (even if no engagement) and police disident opinions

Predict: Model effectiveness

Self-exciting (Hawkes) processes [Hawkes, 1971]



Event intensity

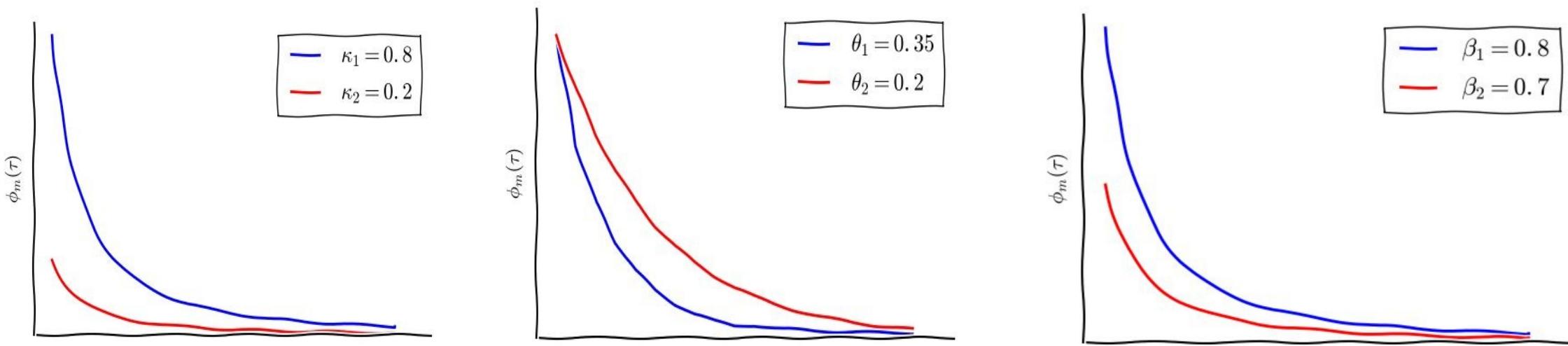
$$\lambda(t|\mathcal{H}_t) = \underline{\mu(t)} + \overline{\sum_{i:t>T_i} \phi(t - T_i)}$$

base intensity
(exogenous)

self-excitation
(endogenous)

Predict: The “social media” kernel

$$\phi_m(t - T_i) = \frac{\kappa m^\beta (t - T_i)^{-(1+\theta)}}{\text{the rate of 'daughter' events} \times \text{content virality} \times \text{user influence} \times \text{memory}}$$



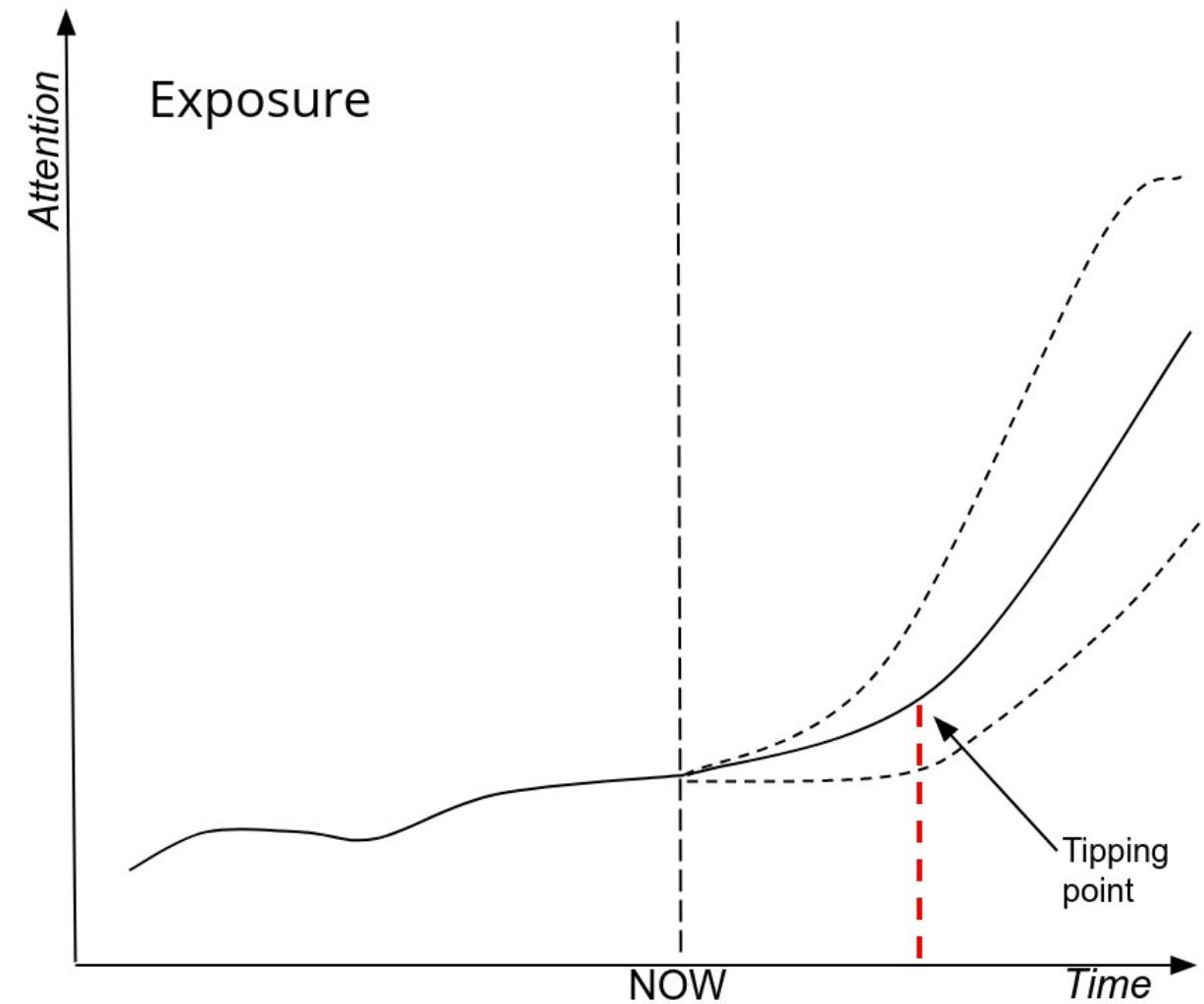
Predict: Future adoption and tipping points

Given an early observed spread, predict the future attention a narrative will gain

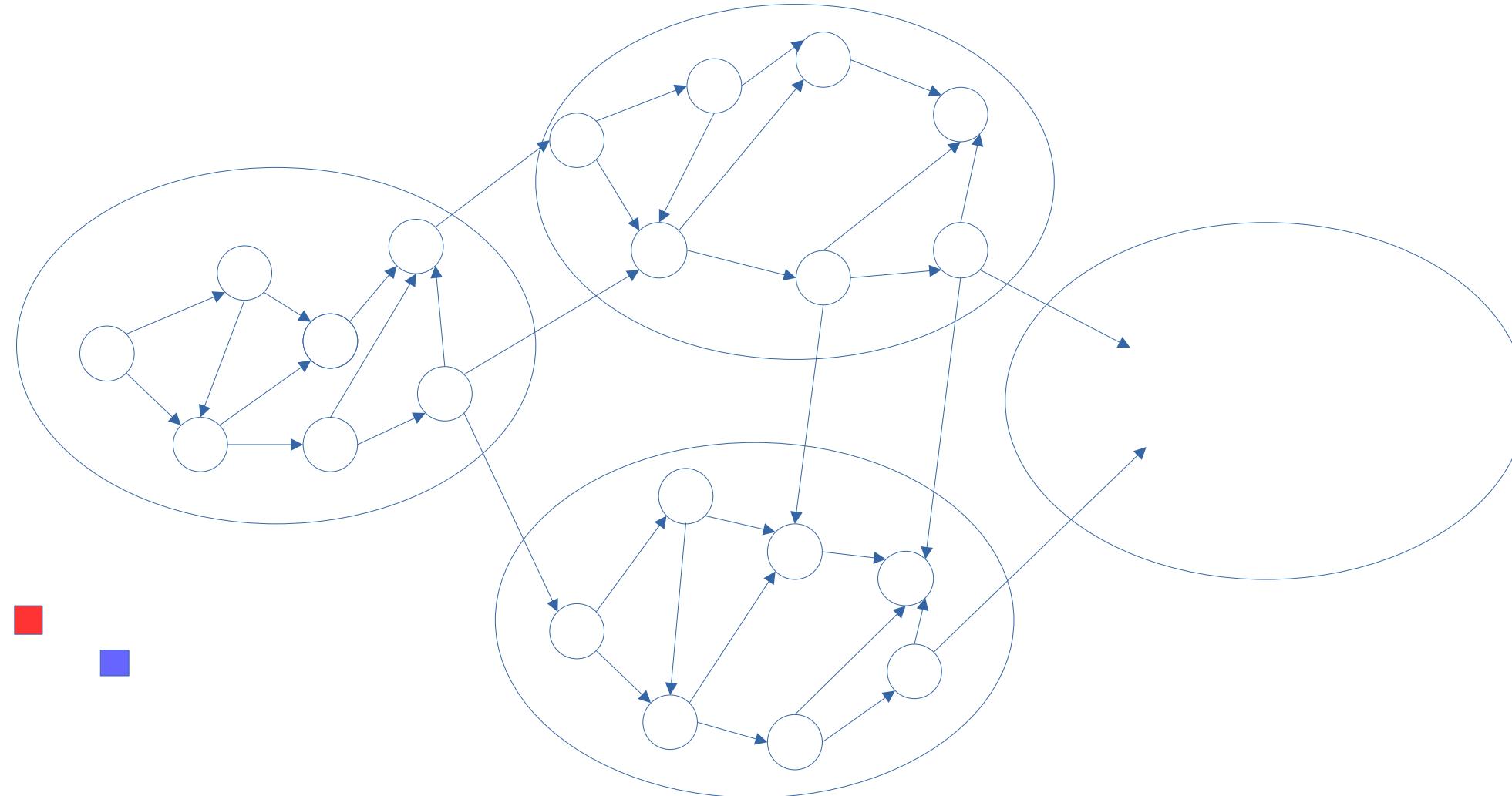
tipping points – the points at which the reach of IO increases exponentially

When to act? (**hint:** before tipping point)

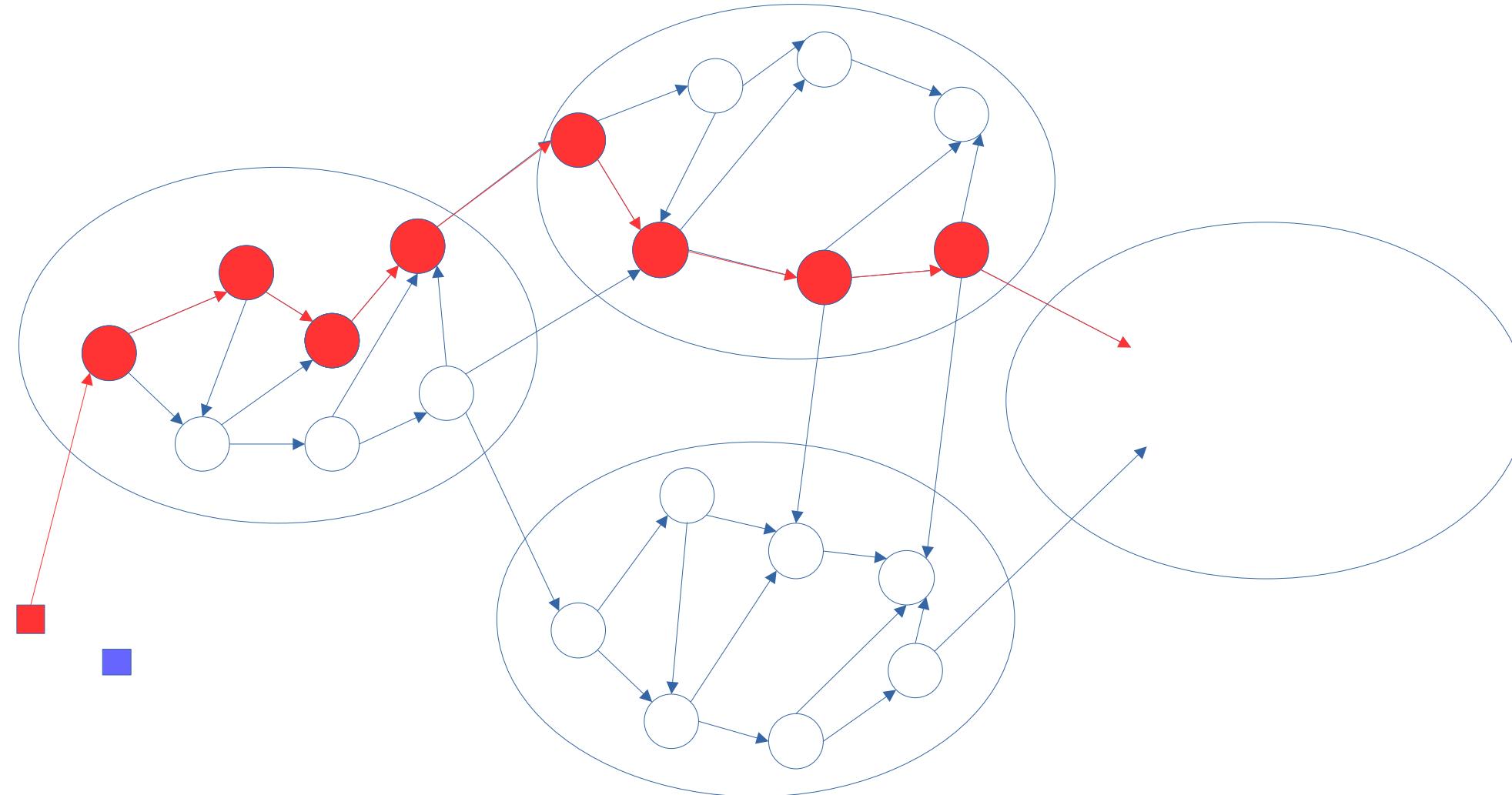
What to counter? (**hint:** potentially popular IO)



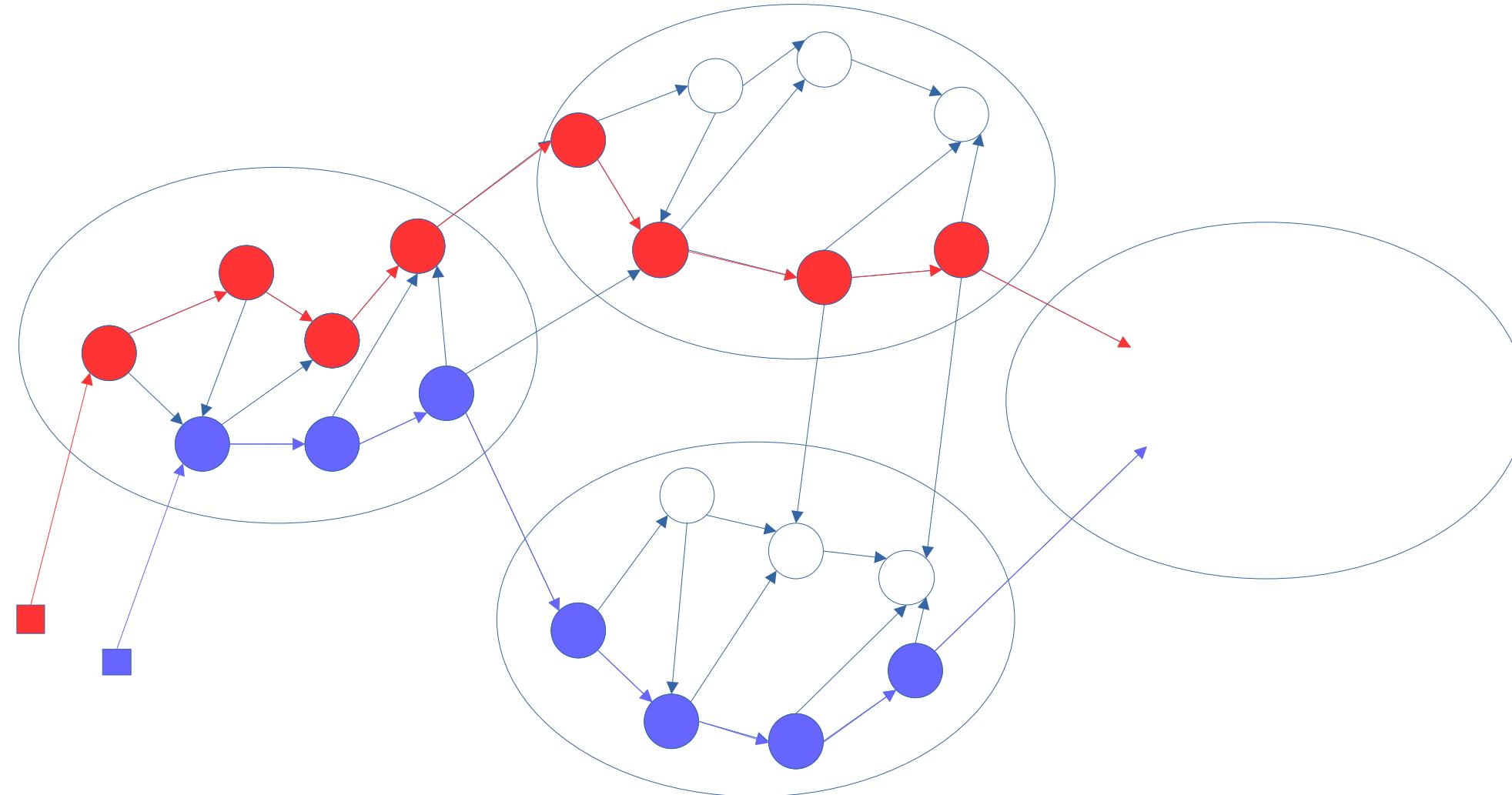
Predict: social neighborhoods on the “path” of a spread



Predict: social neighborhoods on the “path” of a spread



Predict: social neighborhoods on the “path” of a spread

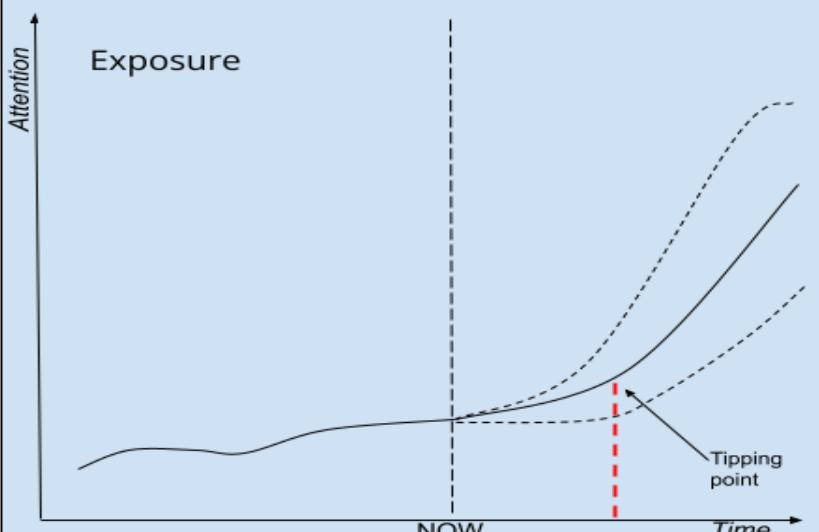


Predict: Effectiveness of state-sponsored narratives

IO Dashboard

IO Episodes

Exposure



Attention

NOW

Time

Tipping point

Content stream

-
-
-
- X
-
-
- X
-
-
- X

<Dashboard console >

Messaging

- Topic 1
- Topic 2
- Topic 3
- Topic 4
- ...
- ...

Most active users

	Username # followers # postings Network centrality #
	Username # followers # postings Network centrality #
	Username # followers # postings Network centrality #

Action plan



**Response
level**



Objective



Approach

Monitor

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Mitigate: countering conspiracy theories, misinformation and state-sponsored IO



Questions:

RQ1: Are messages written in particular styles more effective in spreading?

RQ2: Do different audiences/segments of the population prefer specific styles?

RQ3: Can we use Generative Language models (ChatGPT-like) to construct effective interventions “in the style of...”

RQ4: Can we test the interventions in simulated environment before deploying them?



Prof Hany Farid
UC Berkeley

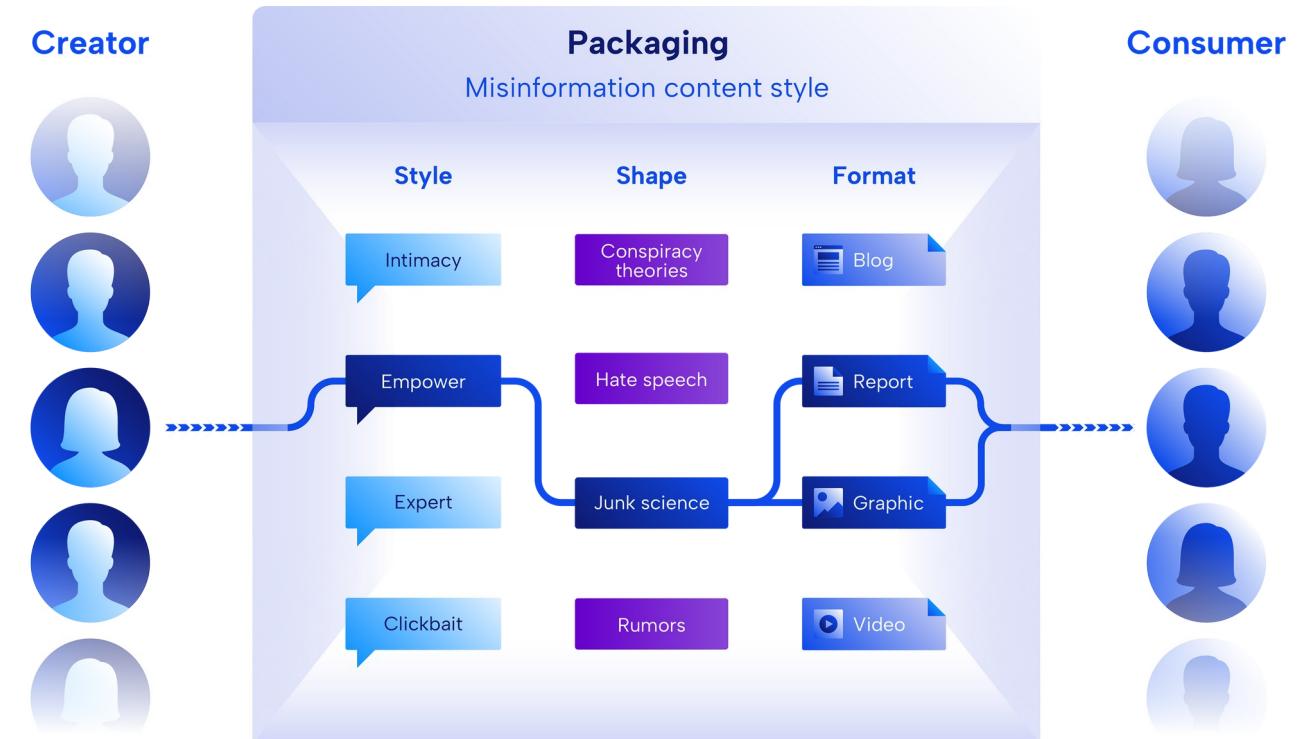
Mitigate: misinformation as a content consumption problem

1

Aim: understand how misinformation spreads in Australia's online ecosystem.

Findings:

- Misinformation is an online product targeting specific vulnerable populations.
- People believe and share misinformation that aligns with their preexisting beliefs and communication style.
- Tailored messaging is 50-100% more effective than generic communication.
- Identify vulnerable users based on their communication style rather than content.



Vulnerable target personae – Jennifer

Jennifer Matthews

age: 25-40

residence: Australia

education: BA

occupation: Formerly a communications specialist for a non-profit, but retired after the birth of her first child

marital status: Married with two children (a newborn and a 2 year old)



"A mother's work is never done"

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and painting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life and it's not very interesting to him.

Interests:

Health, children, creativity, simple pleasures

Fears

Her children will get sick
She is not the best mother she could be
Her friends at work forget her
She become unattractive to her husband after having two children

Preferred style:

Intimacy + Junk Science

Barrier to generic communication:

formal, masculine style

Vulnerable target personae – Patrick

Patrick Bear

age: 67

residence: Suburban NSW

education: Bachelor of Education

occupation: Retired primary school teacher

marital status: Married for 45 years, 3 children, 4 grandchildren



"Kids these days"

Patrick wakes early and takes a walk around the neighbourhood, before helping his wife up and organising breakfast for her. She's become quite frail, and he has to help her get dressed and walk. After, they often sit outside in the garden and he reads to her. He organises lunch and they both have a short rest, before the grandkids come over after school until their parents finish work. These frantic few hours are their favourite part of the day, even if it gives them a bit of a headache. He helps the kids with homework, and then takes care of chores as his wife supervises them watching TV or spending time in the garden. When the grandkids have been retrieved by their parents he organises dinner, and as his wife watches late night TV, he logs on to the Facebook his daughter signed him up for. He tries to keep up with the posts there, and also finds interesting news items about issues he never hears about on TV. His wife thinks the internet is silly, but he's excited about the new things he can learn there. As he gets ready for bed, he often finds himself still thinking about the things he's read.

Interests:

Humour, Hard work, Family

Fears

New changes in the world

His family forgetting him

Financial security

Being alone

Illness and death for him and his wife

Preferred style:

Expert + Conspiracy

Barrier to generic communication:
belief in own expertise, difficulty in understanding long complex text

Vulnerable target personae – Aaron

Aaron Turner

age: 21

residence: Australia

education: Law student (final year)

occupation: Summer law clerk

marital status: Single



"I object!"

Aaron wakes at around 8am and makes a mad rush for either university or his summer job, skipping breakfast. He is enthusiastic about both and works hard at each. By lunch time, he is feeling exhausted and deserving of a "treat", often buying a burger or pizza meal--and after all, he skipped breakfast, which makes up for the extra calories. He works intently through the afternoon, and leaves work feeling exhausted again. He often tells himself that he's too tired to hit the gym now, but he'll go tomorrow. On Fridays, he spends the night with his Church youth group, but it's increasingly bittersweet, as he will have to leave it after his upcoming birthday. On nights when he doesn't have Church, he often winds up having a few beers and scrolling through Reddit, where he can vent his frustration about still being single and find other interesting blogs to explore.

Interests:

Strength and confidence in men

Beauty in women

Christianity

Recognition for his hard work

Fears

Professional failure

Personal failure – He will not get a girlfriend

He will lose his connection to his Church

Preferred style:

Empower + Hate Speech

Barrier to generic communication:

authoritative tone

Mitigate: run “packaged” ads campaigns



Instruction set:

Write a 100 word social media post that would appeal to Jennifer, accounting for:

- the post comes from a fellow mother who understands Jennifer's fears
- she reassures her that vaccines are safe for her children.
- she vaccinated her children against smallpox the other day, and 2 weeks later all the children in the neighborhood got it, except hers.
- write it as if it is send on a mum's Facebook group.

A synthetically-generated social media post , campaigning for vaccines targeting the “Jennifer” persona.

Mitigate: run “packaged” ads campaigns



Generated post:

Hey mamas,

I just wanted to share something with you all that happened recently. I vaccinated my kids against smallpox a couple of weeks ago and, unfortunately, all the kids in our neighborhood got it except mine. It's a scary thing when there's a new vaccine out there, but I just wanted to let you all know that vaccines are safe and effective. They've been tested a ton, and they work to protect our kids from some really dangerous diseases. So if you're still on the fence about vaccinating your kiddos, I'm here to tell you - do it!

Let's keep our babies healthy and safe together!

#MomSquad #VaccineSafety #SmallpoxPrevention
#HealthyKids.

Love,
One of your fellow moms.

A synthetically-generated social media post , campaigning for vaccines targeting the “Jennifer” persona.

Mitigate: Effectiveness of communication styling

Testing protocol:

- Facebook ads with A/B testing facility
- Stock images
- Ads link to Government information sites
- A/B test: targeted (A) vs generic styling (B). All other indicators stay the same: target demographic, image, message

Tested variables:

- Three target demographics – consumer personae
- Four message styling – generic, Jennifer-, Patrick- and Aaron-styled
- 5 topics of interest: children vaccination, children sexual education, 5G, climate change, nutrition supplements for physical fitness

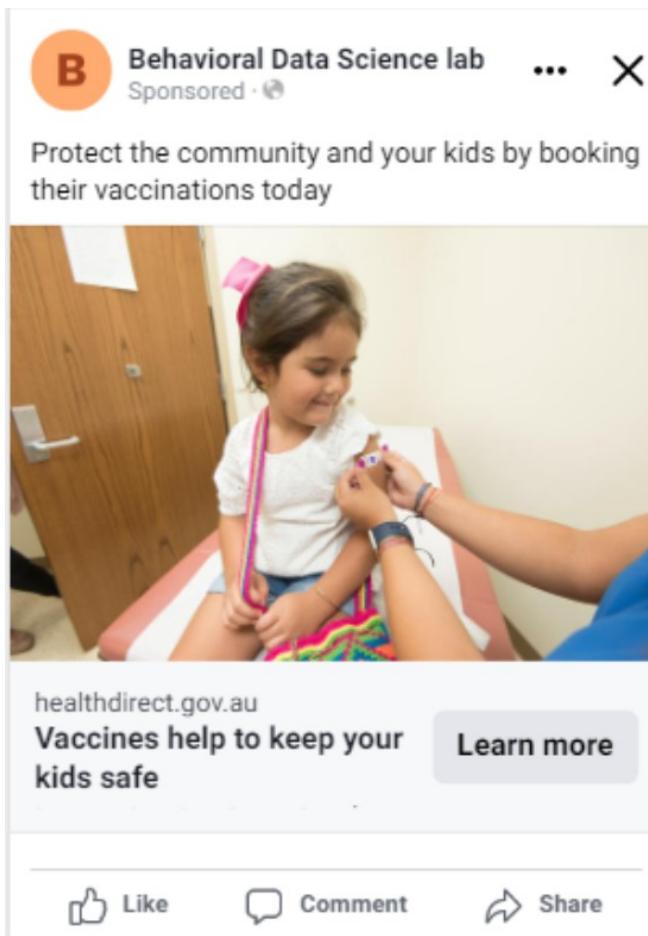
Success measurement

- Click-Through-Rate (CTR): number of clicks vs number of impressions
- **Higher (CTR) – higher message effectiveness**

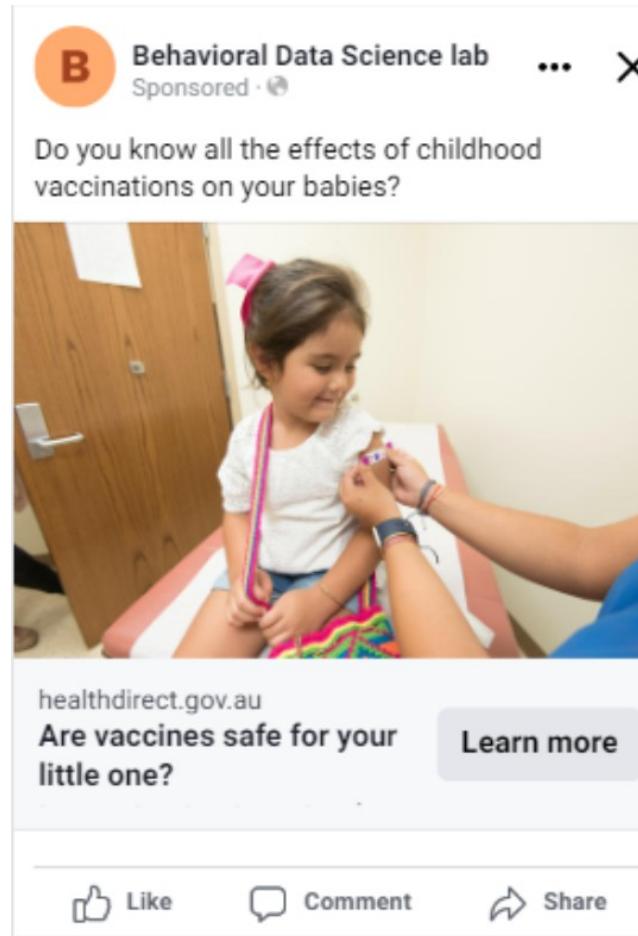
Mitigate: Effectiveness of communication styling

Children vaccination (Jennifer)

A: Generic



B: Styled



Results:

A: Generic

Reach: 10,346 users

Impressions: 15,567

Clicks: 101

CTR: 0.65%

B: Styled

Reach: 10,020 users

Impressions: 13,107

Clicks: 129

CTR: 0.98%

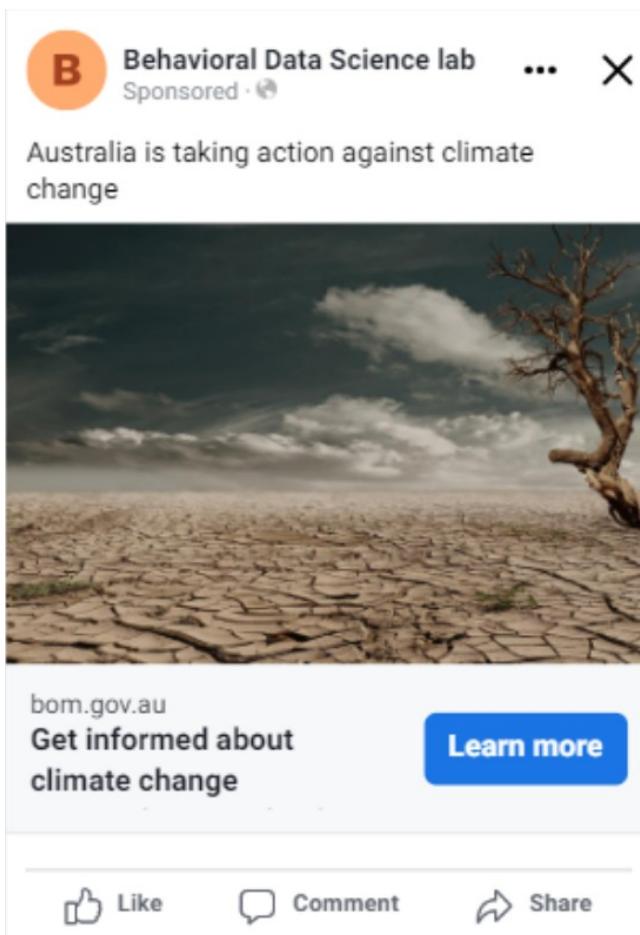
Improvement:

51%

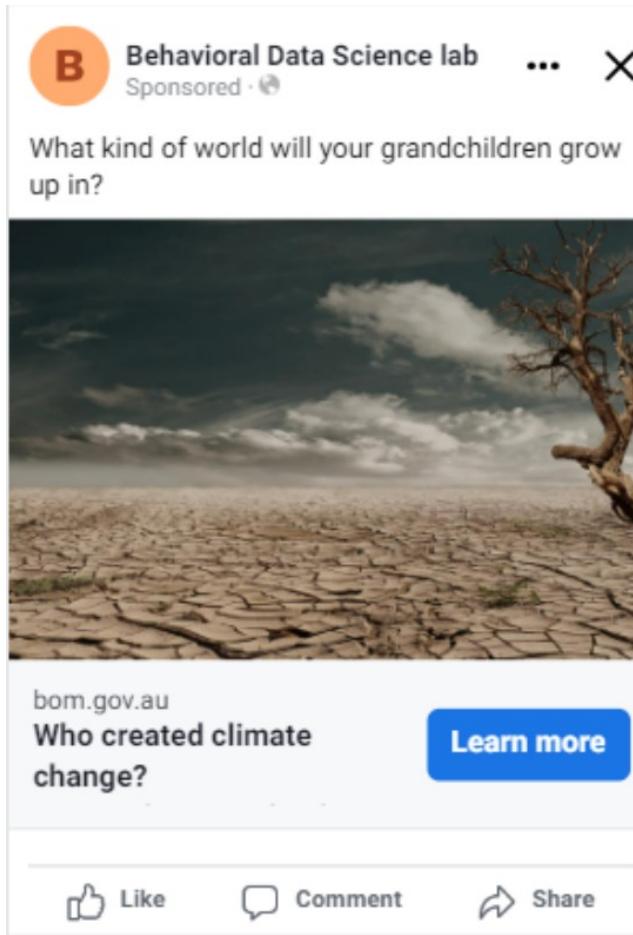
Mitigate: Effectiveness of communication styling

Climate change (Patrick)

A. Generic



B. Styled



Results:

A: Generic

Reach: 5,362 users

Impressions: 10,095

Clicks: 170

CTR: 1.68%

B: Styled

Reach: 5,952 users

Impressions: 10,082

Clicks: 334

CTR: 3.31%

Improvement:

97%

Action plan in Countering Foreign interference and state backed IO



**Response
level**



Objective



Approach

Monitor

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

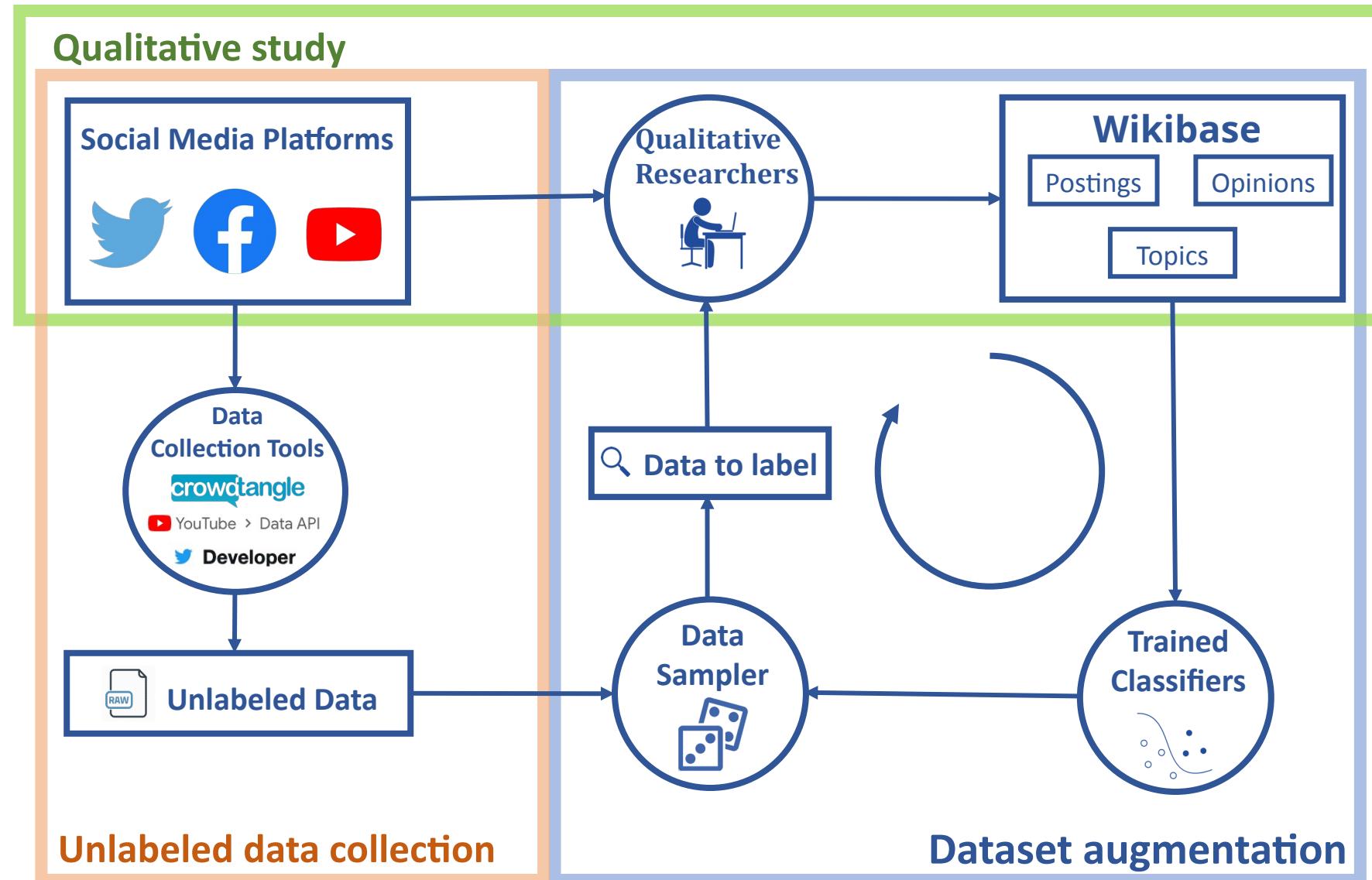
What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

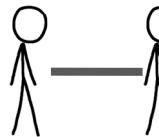
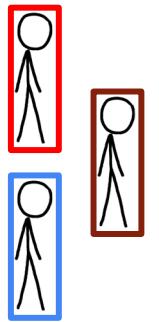
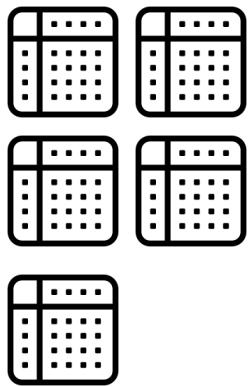
Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

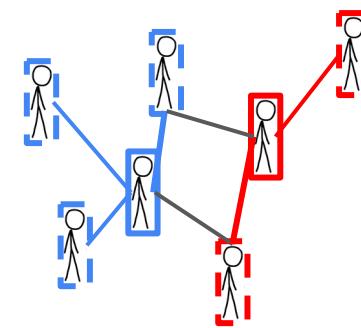
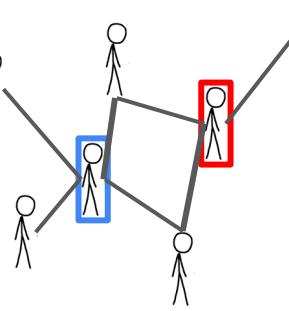
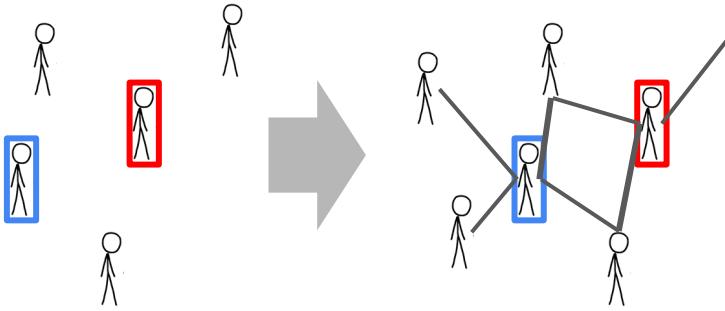
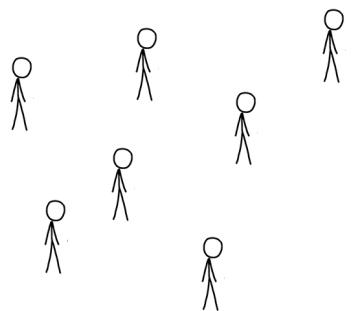
Monitor: Semi-automatic data labelling



Detect: far-right and extremist ideology detection



 LightGBM



No data labelling required for new datasets

Seed users: use fringe news sharing to label a small subset of users

Text-based homophily quantification – “what you write is who you are”

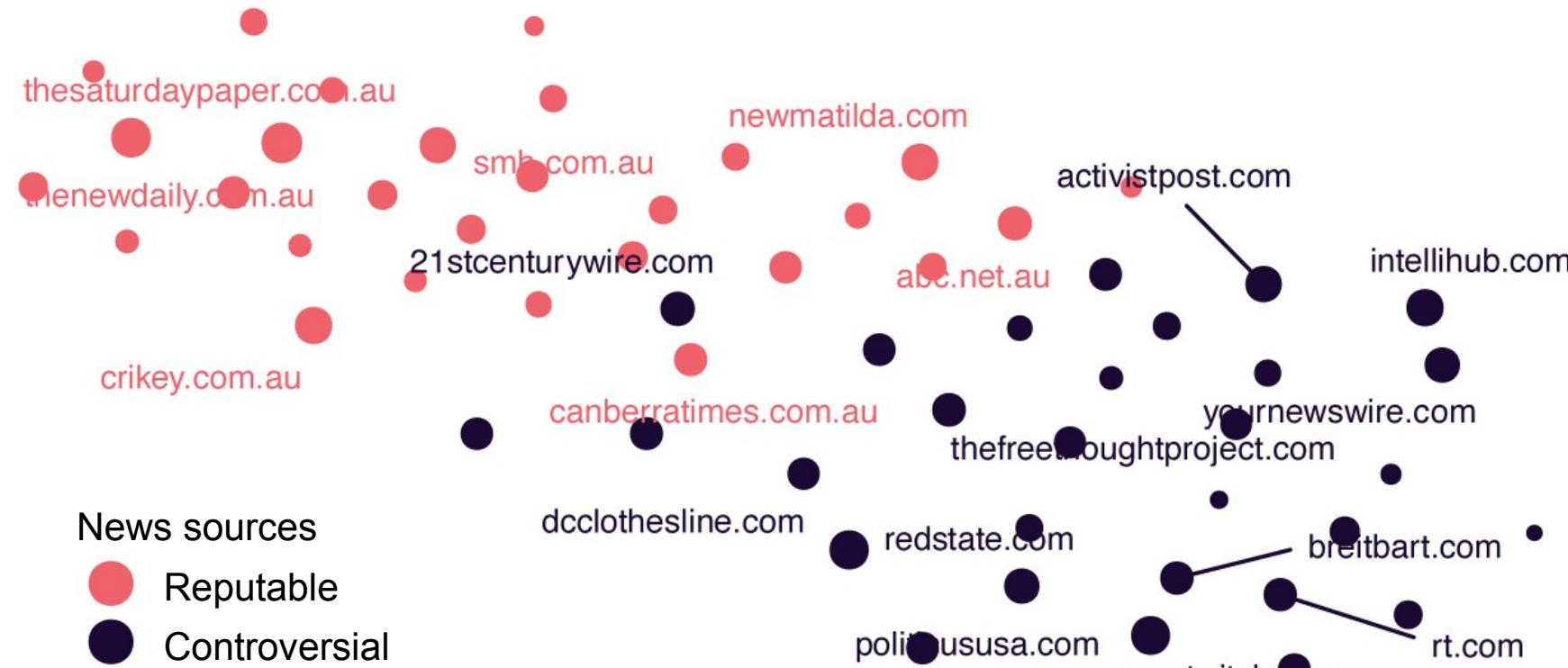
Use advanced Machine Learning to label all users



The technical detail:

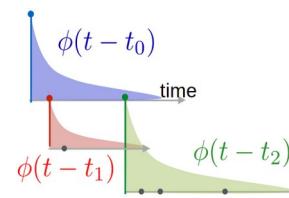
Ideology proxies; homophily lenses (text, follower, URLs); automatic user labelling

Detect: separating controversial from reputable



Reputable and controversial sources are separable based solely on how their information spreads

Detect controversial news without content analysis



evently

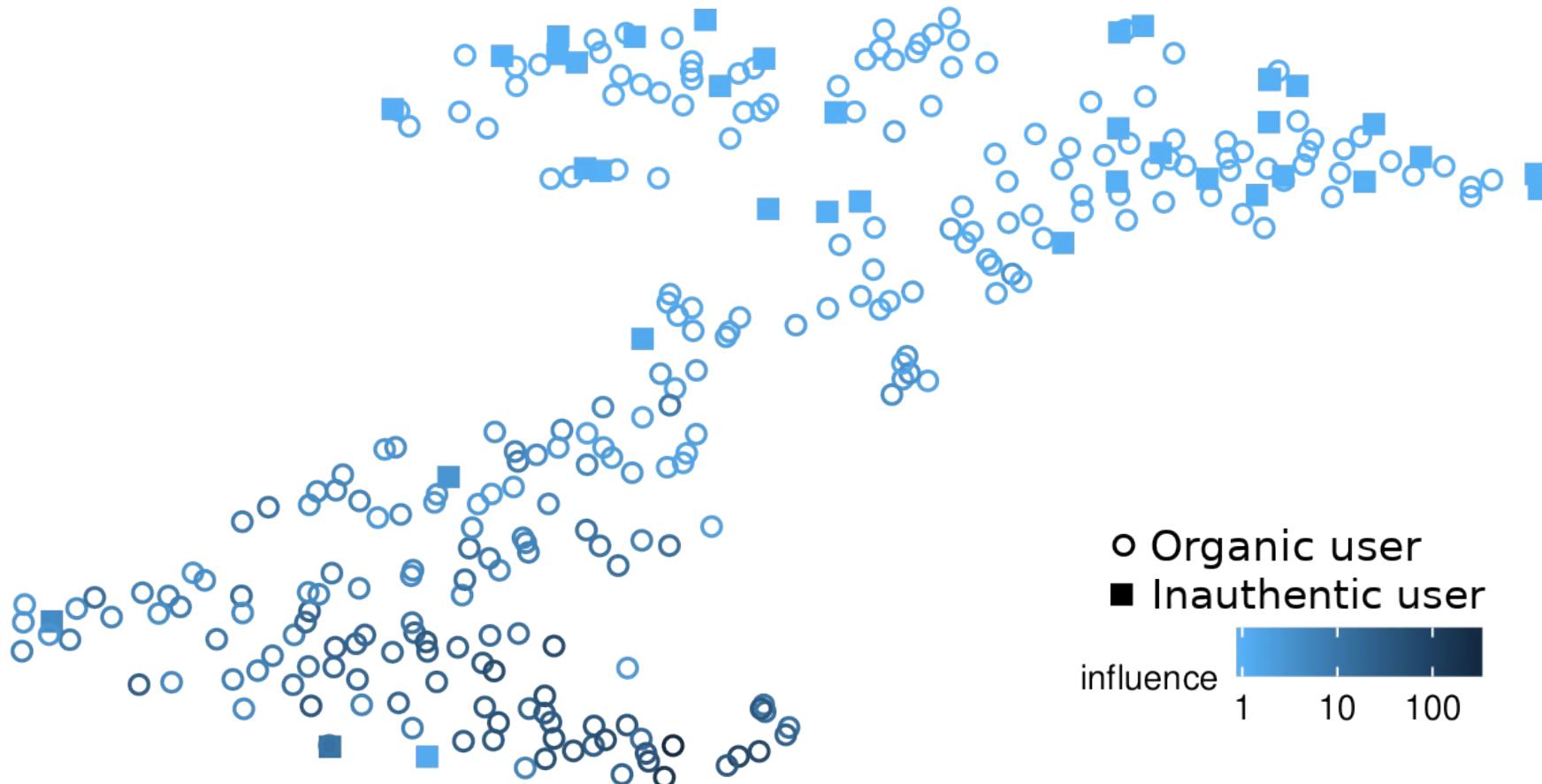
https://www.behavioral-ds.science/theme1_content/evently/



The technical detail:

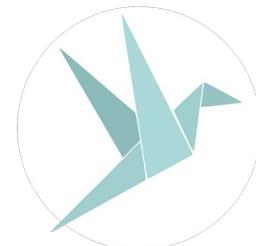
Mathematical generative modelling; Hawkes processes; joint modelling

React: Identify influential inauthentic users (bots)



Identify users engaged in influence operations

Estimate their impact on the wider community



birdspotter

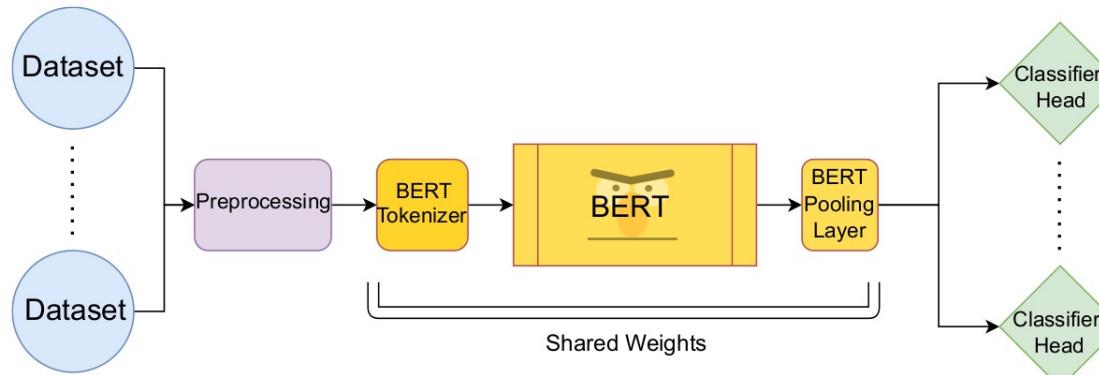
https://www.behavioral-ds.science/theme2_content/birdspotter/



The technical detail:

Influence estimation using stochastic modelling; content-free analysis

React: Detecting Hate Speech in Unseen Domains



Novel learning paradigm to leverage many disparate datasets to learn a single hate speech representation

Improved performances over the state-of-the-art, generalizable to novel datasets.

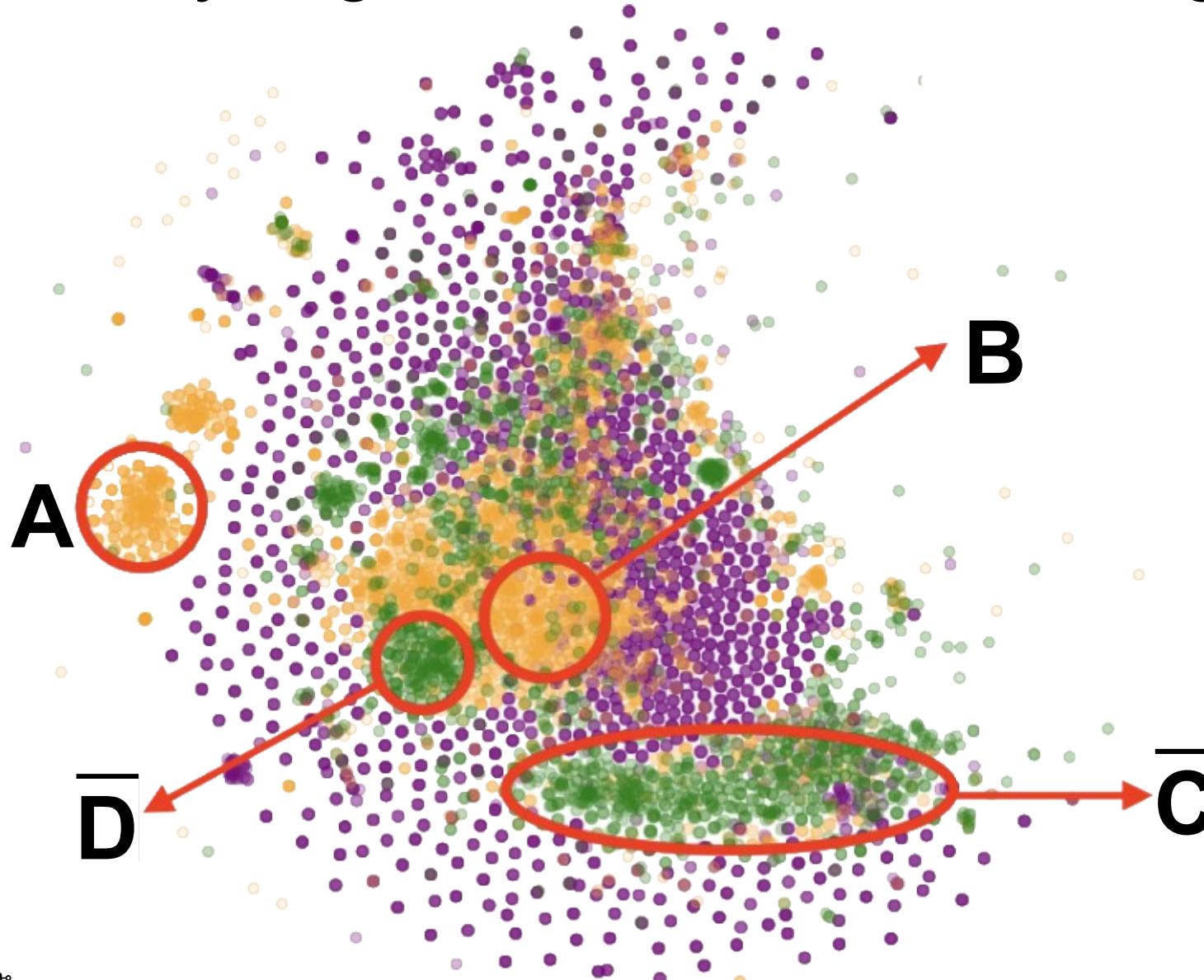
Model		Testing Dataset									# Wins
		DAVIDSON	WASEEM	REDDIT	GAB	FOX	STORM- FRONT	MANDL	HATEVAL	PUBFIGS-L	
MTL	MTL-NCH	0.6822	0.3801	0.8456	0.8738	0.6150	0.6826	0.5312	0.6449	0.6175	6
	MTL-MV	0.6455	0.4048	0.8263	0.8660	0.6030	0.6771	0.4834	0.6315	0.6231	1
BERT baseline trained on:	DAVIDSON		0.5556	0.5914	0.6731	0.4932	0.4597	0.5690	0.5414	0.5469	0
	WASEEM	0.6136		0.6000	0.6427	0.5519	0.5356	0.5099	0.5784	0.5611	0
	REDDIT	0.6135	0.4957		0.8083	0.5229	0.5559	0.4900	0.5741	0.5402	0
	GAB	0.5720	0.4595	0.8375		0.5075	0.5645	0.4277	0.5664	0.5185	0
	FOX	0.4285	0.4249	0.4234	0.4651		0.3865	0.4159	0.4490	0.3926	0
	STORMFRONT	0.4533	0.5467	0.5822	0.6487	0.5740		0.5104	0.5664	0.5659	0
	MANDL	0.3336	0.4822	0.4066	0.4582	0.4010	0.3518		0.4546	0.3633	0
	HATEVAL	0.5849	0.5824	0.5700	0.5796	0.5532	0.5466	0.5348		0.5432	0
	PUBFIGS-L	0.6351	0.6048	0.5970	0.6600	0.5546	0.5249	0.5963	0.5858		2



The technical detail:

Transfer learning; language models fine-tuning;

Analysing coordinated troll strategies



(yellow) right trolls: focused MAGA
(magenta) left trolls: surround discussion
(green) news trolls: selective highlighting

A – (right trolls) Hillary cannot be trusted
#ThingsMoreTrustedThanHillary

B – (right trolls) Mimic black Trump supporters
#Blacks4Trump

C – (news trolls) News about violence and civil unrest
#news

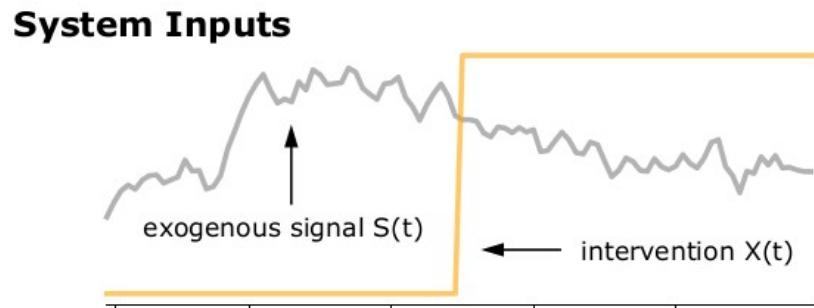
D – (news trolls) Federal politics, policy and regulation
#politics



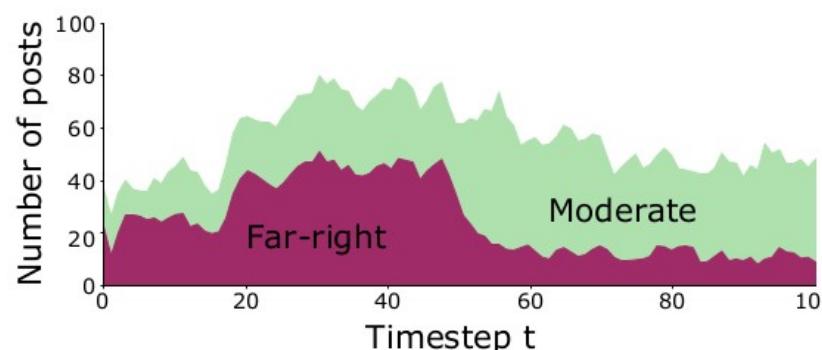
The technical detail:

Semantic edit distance; dimensionality reduction; Twitter trolls

Mitigate: synthetic testing of interventions

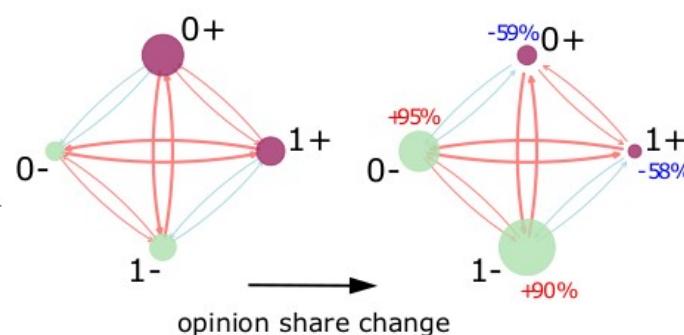


Tier 1
(attention volumes)



Tier 2
(market shares)

○ majority opinion ○ minority opinion



Mitigate: synthetic testing of interventions

