UTS

An online disinformation wildfire

monitoring, detecting and reacting to violent extremism and foreign interference

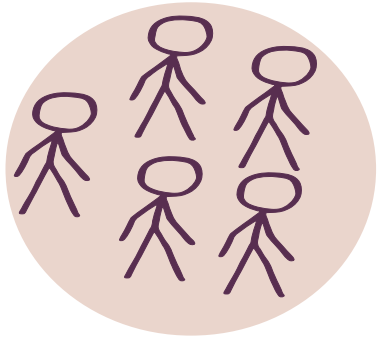
Data Science Institute

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

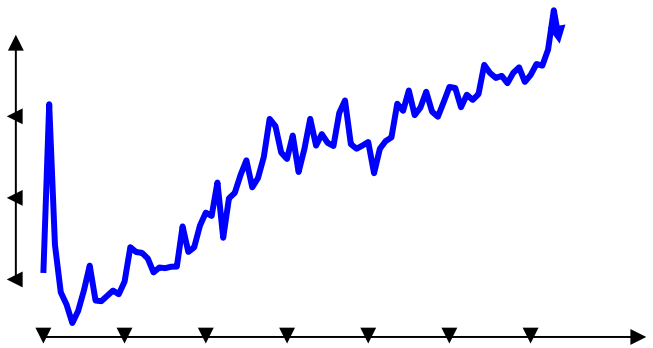
UTS CRICOS 00099F

The Behavioral Data Science

1.

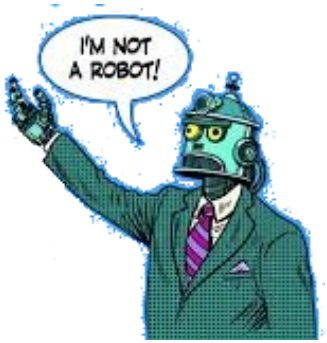


information diffusion
epidemics spreading
behavioral modeling



3.

2.



[Rizoiu et al ICWSM'18]

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

Prior expertise

Australian National University
CRAWFORD SCHOOL OF PUBLIC POLICY

Tracking Disinformation Campaigns across terrain

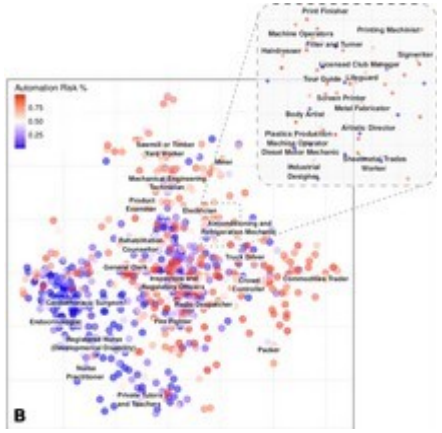
Australian Government
Department of Defence
Defence Science and Technology Group

Real-time detection of disinformation campaigns

Hate Speech propagation on Social Media

NSW GOVERNMENT

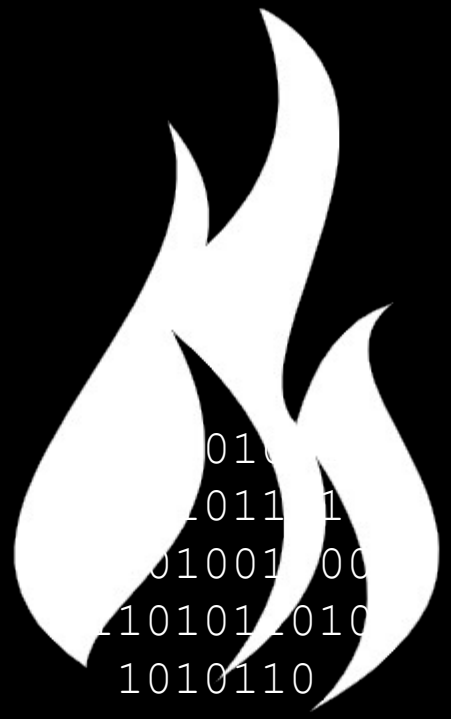
Expert roundtable for Defamation law reform



Occupation transition recommender systems

WIKIMEDIA FOUNDATION
WIKIPEDIA The Free Encyclopedia

Detecting and quantifying privacy loss over time



Disinformation wildfires spreading across Australia

The 2019 Australian elections?

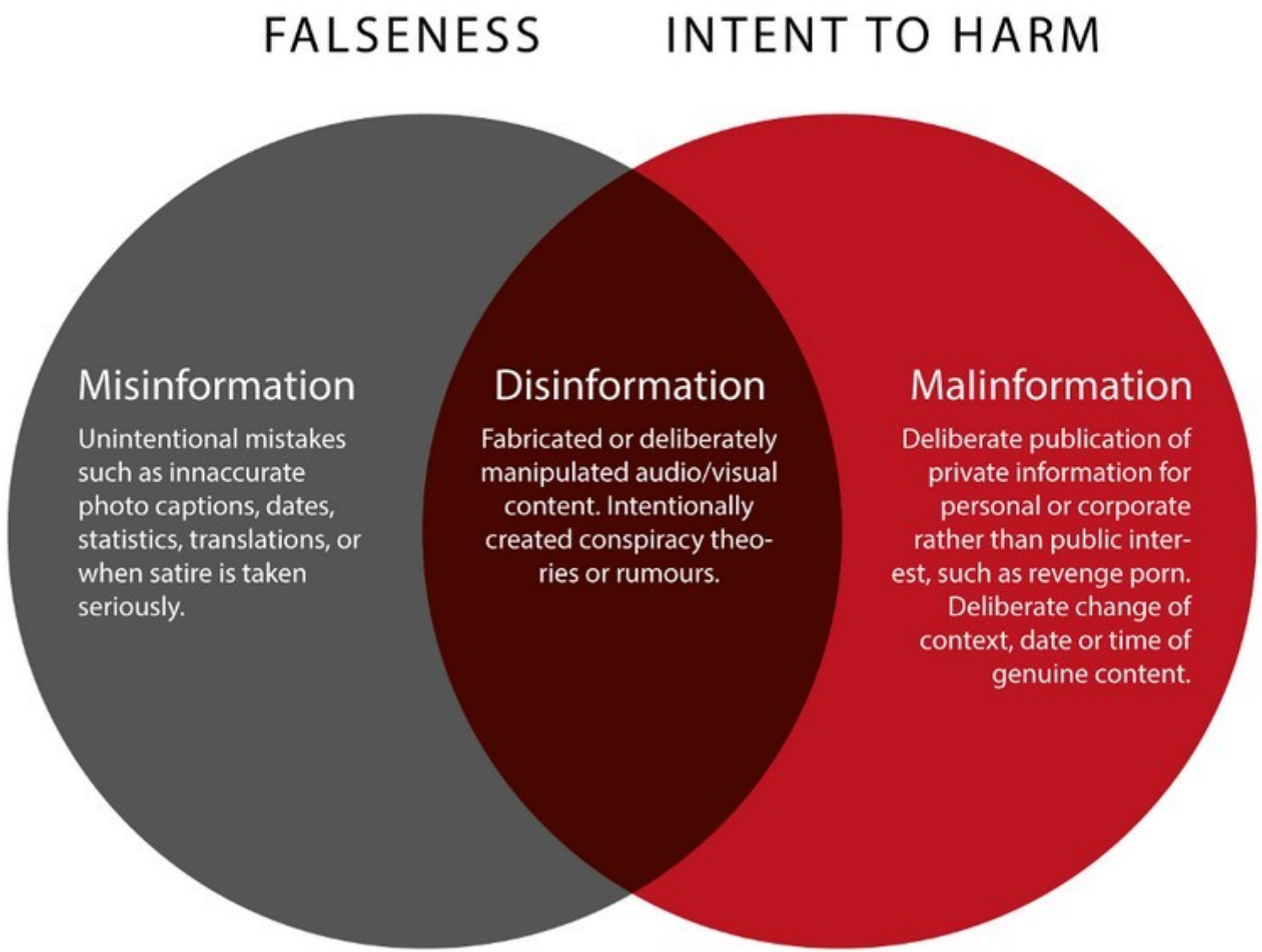
death tax: Labour intends to institute a tax on inheritance

To vote correctly put a number 1 next to Liberal

7NEWS Melbourne - St Kilda gang rampage | Faceb...

St. Kilda is terrorized by African gangs

TYPES OF INFORMATION DISORDER



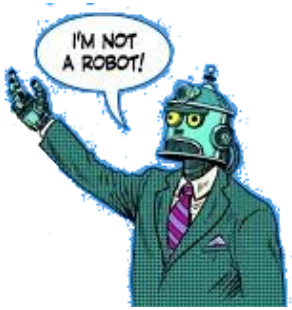
Red Queen effect

Content-based detectors are sensitive to adversarial training attacks – simply use the detector to train the attacker.

Our approach in a nutshell

Use the reaction of the social system for early detection – build detection systems based on the spread patterns of mis-/dis-/information within the user population

Content types and user actions can be distinguished based on how online social systems react to them.



Countering influence and disinformation campaigns – **expertise:**

A. Detecting coordinated campaigns

B. Identify influential inauthentic users

C. Analysing coordinated troll strategies

D. Separating controversial from reputable

A. Expertise: Detecting coordinated campaigns

Clear structure with two clusters: disinformation (right) and debunking (left)

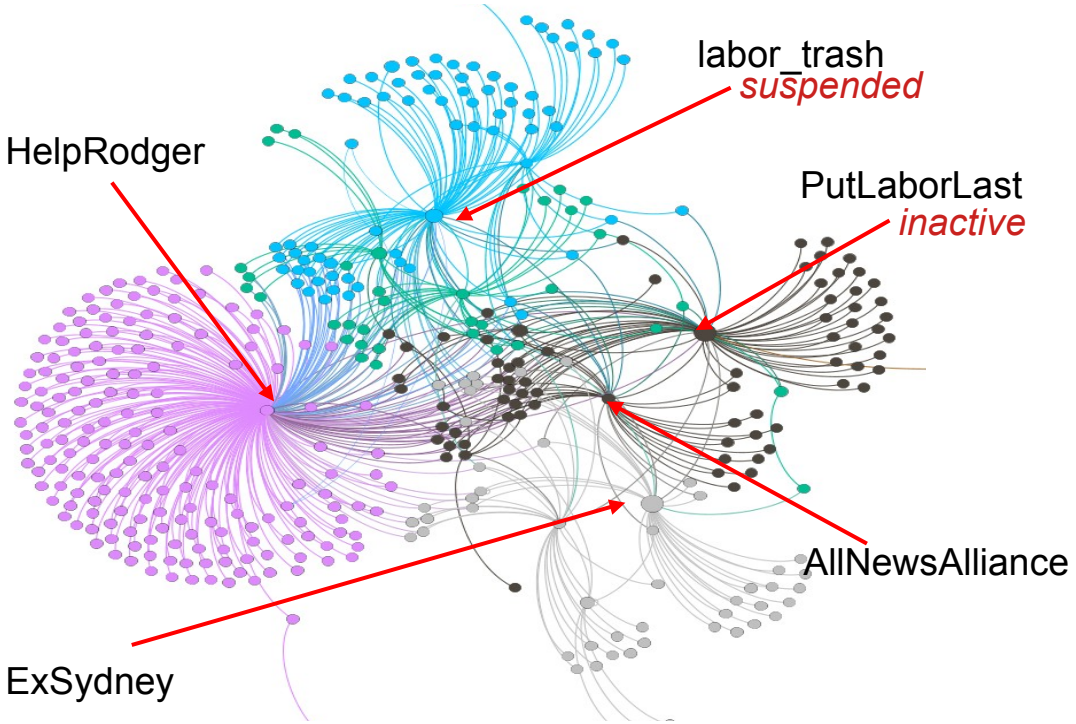
Disinformation cluster: tightly connected, coordinated and timed retweeting

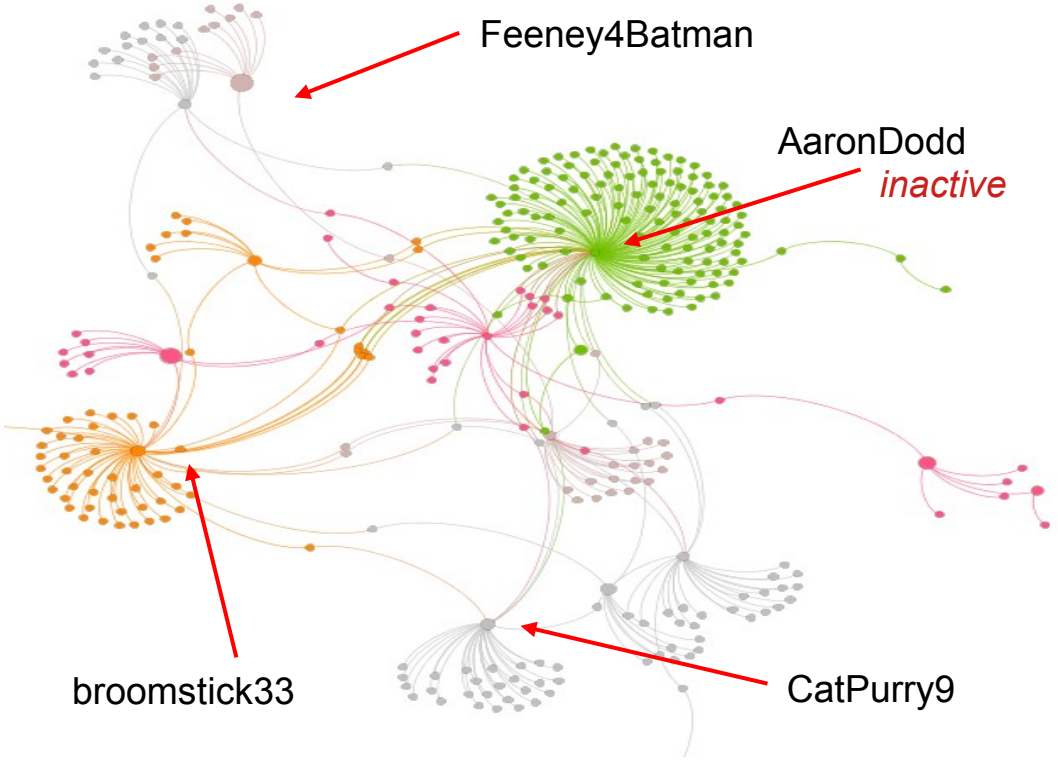
Debunking cluster: organic retweeting, reactionary, loosely connected, multiple communities

The technical detail:

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

Two clusters emerge





Observations:

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

Analyzing narratives (1)

Misinformation cluster

☻ Ramjet ☻
@HelpRodger

#BREAKING
Tax expert has verified that a Shorten Labor party #DEATHTAX means all valuable items such as gold teeth , sentimental jewellery, anything of value will be taxed 40% on death.
I cant believe Labor would do this to families. #auspol #qanda #insiders #9Today #sunrise

☻ Ramjet ☻
@HelpRodger

#BREAKING
Australians are finally aware of Bill Shortens #DeathTax
There is nothing more abhorrent than a Labor govt robbing dead ppls graves.
I'm sick to the stomach thinking about this.#auspol #qanda #insiders #730Report #9Today #sunrise #TheProjectTV #MKR #livingroom #60mins

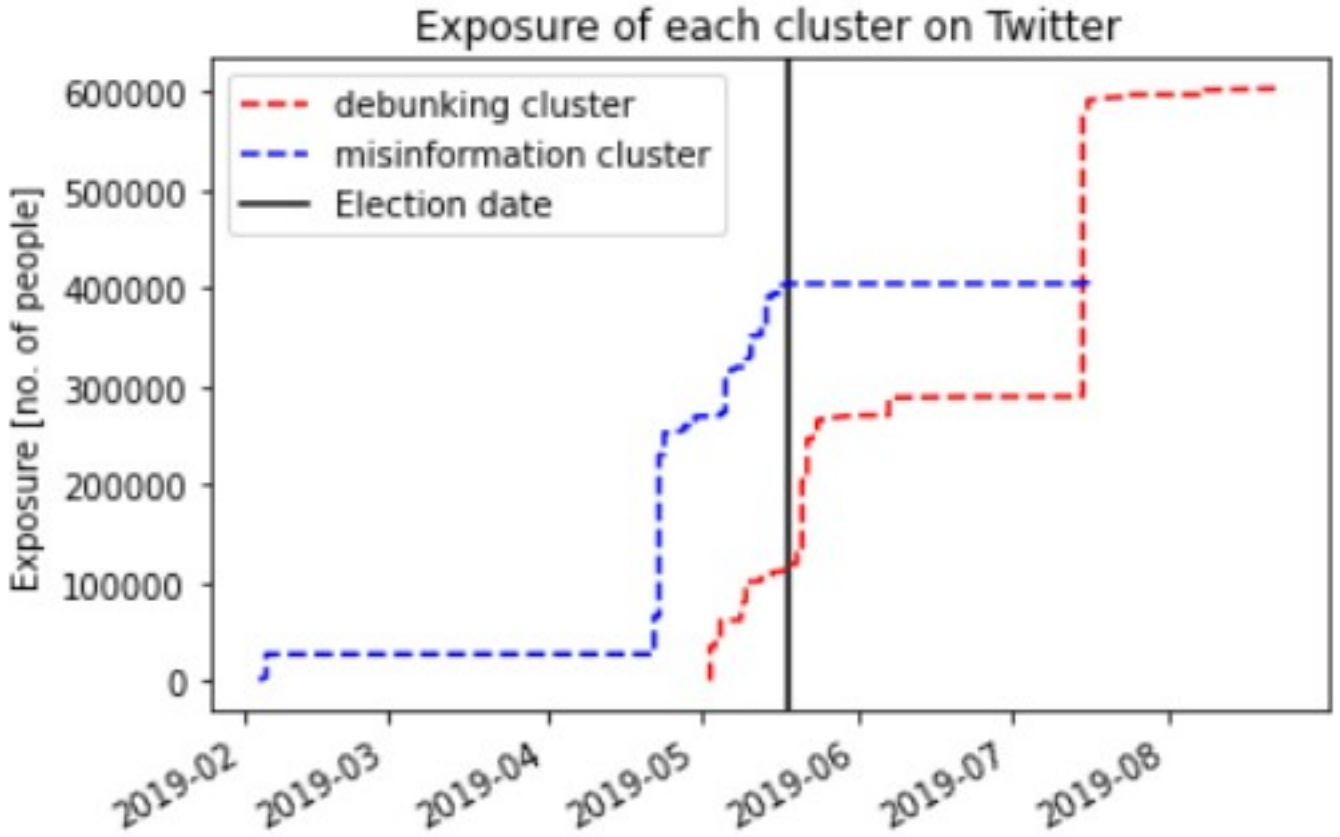
- Perpetuating the misinformation
- Nazi references – “sentimental jewellery”, “gold teeth”, etc.
- Spiteful phrases – “lose parents”, “worth dead” and “truly screwed”
- Confirmative language – “experts verified”

Analyzing narratives (2)

Debunking cluster

- Based around debunking the #DEATHTAX myth
- Phrases like “fabricated”, “completely false”, “scare campaign”

Exposure analysis



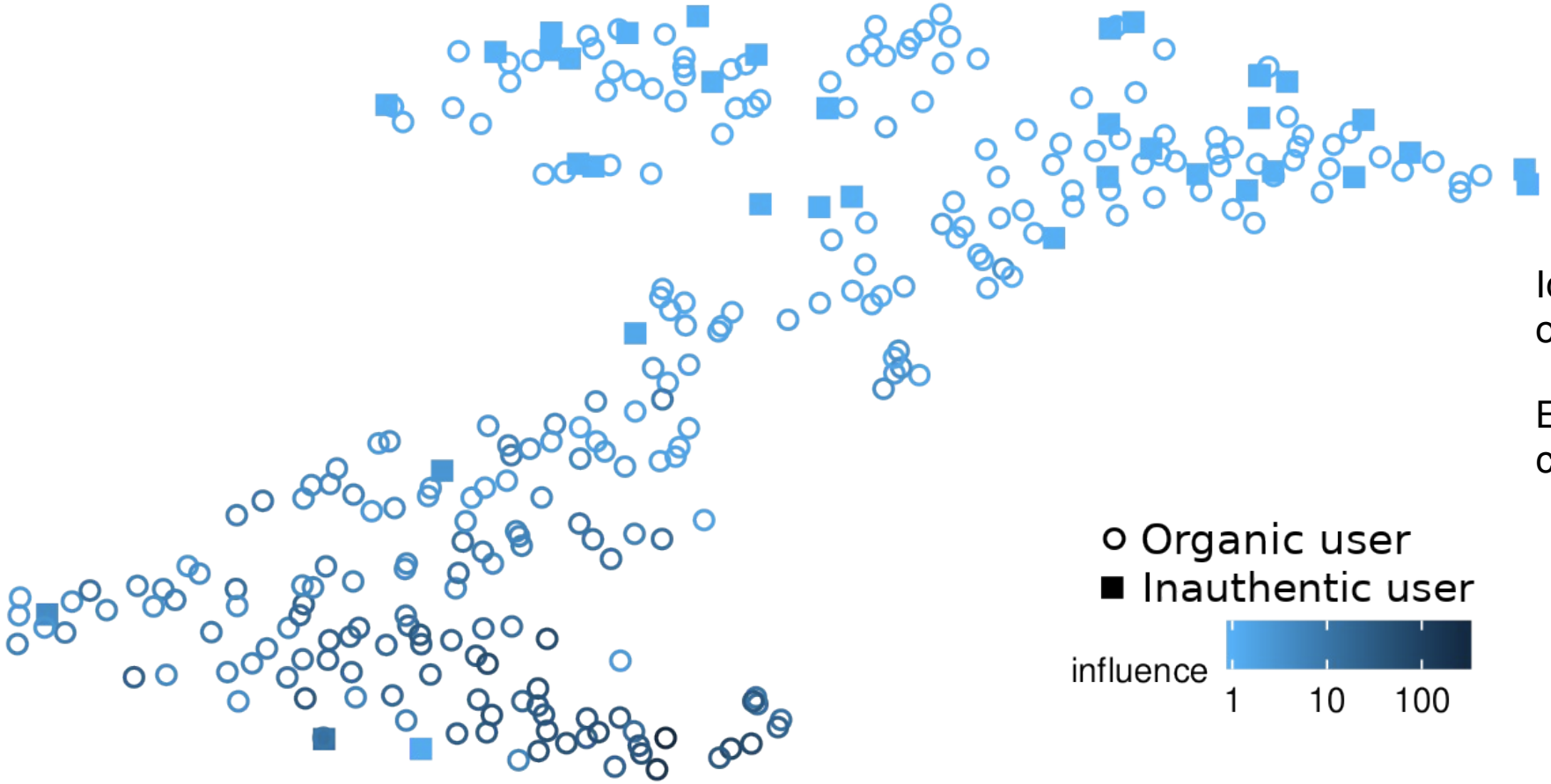
- Early spread of the misinformation campaign
- Exposure of misinformation cluster flattens subsequent to election date
- Debunking cluster gains traction after the election

How Twitter reacted?

	Misinformation Cluster (%)	Debunking Cluster (%)
Suspended Accounts	15 %	2.7 %
Deactivated Accounts	6.2 %	3.1 %
Total	**21.2 %**	**5.8 %**

Nearly ¼ of the users in the misinformation cluster are currently not active (suspended or deactivated).

B. Expertise: Identify influential inauthentic users



Identify users engaged in influence operations

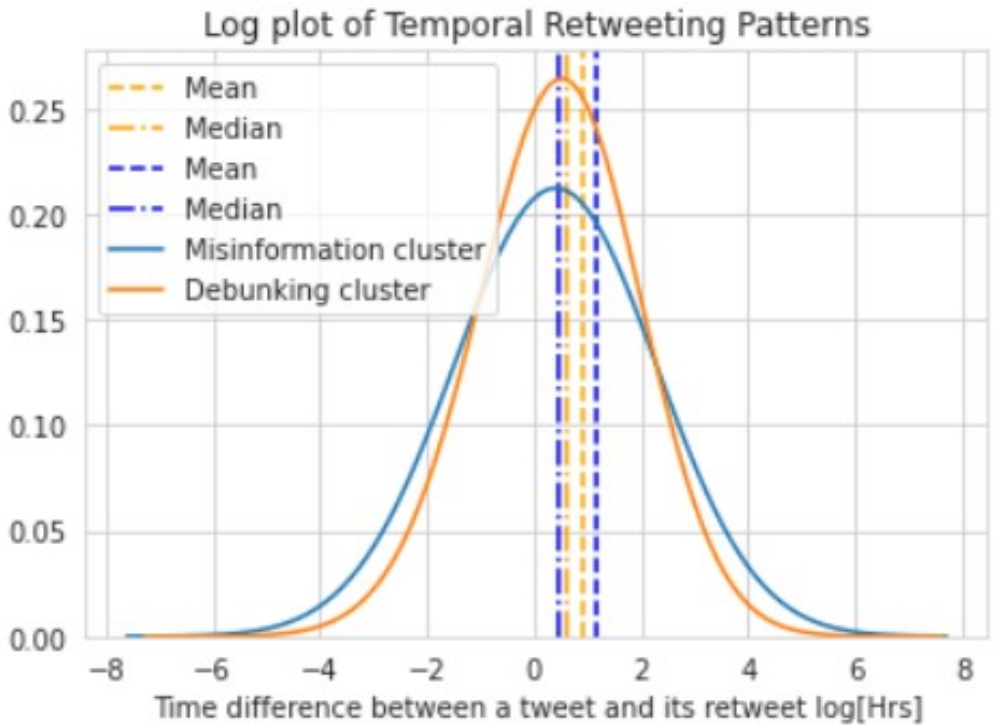
Estimate their impact on the wider community

The technical detail:

Influence estimation using stochastic modelling; content-free analysis

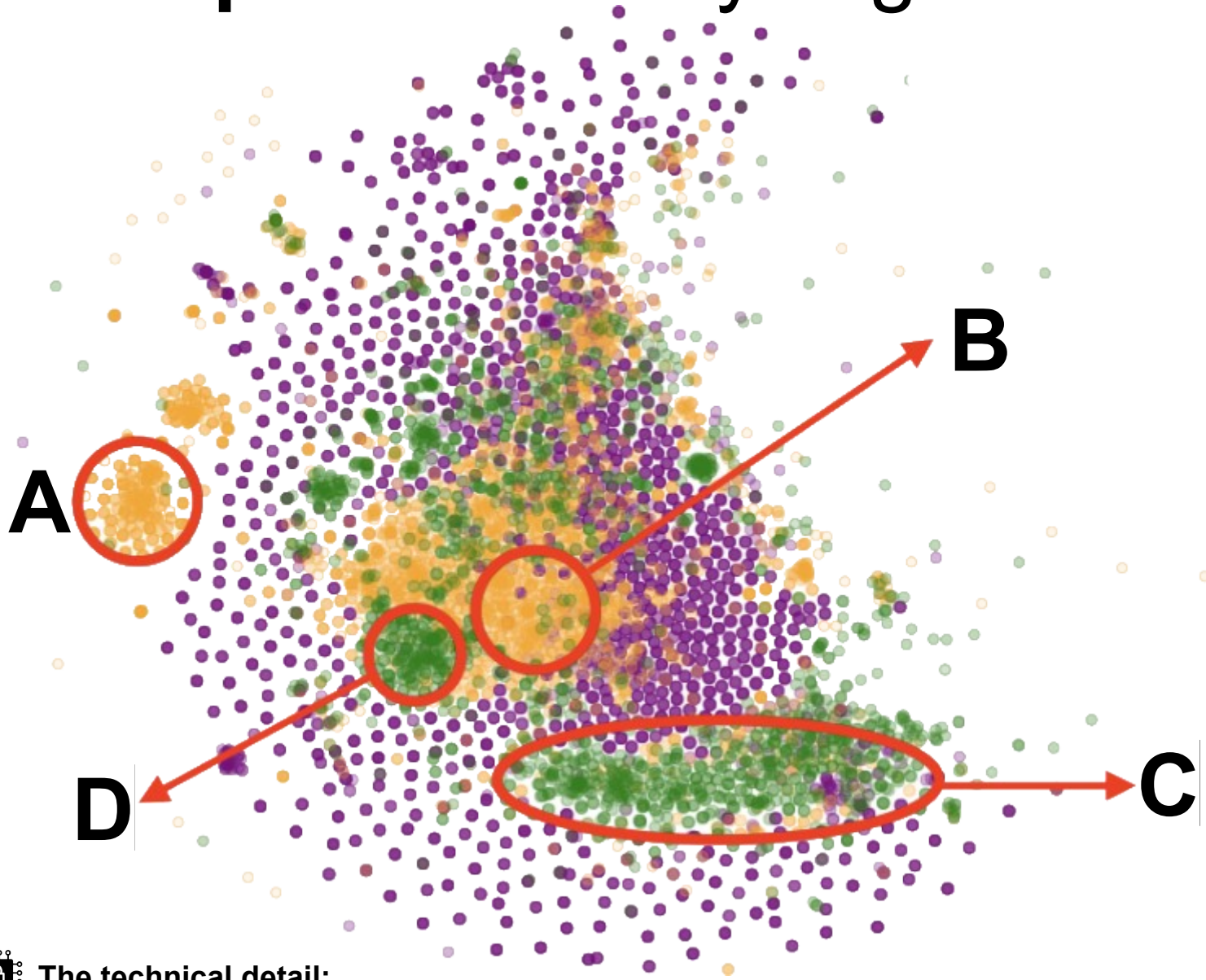
birdspotter

Retweeting patterns for #deathtax users



Misinformation cluster have higher probability of retweeting very quickly (seconds) and very late (days) compared to the debunking cluster.

C. Expertise: 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*

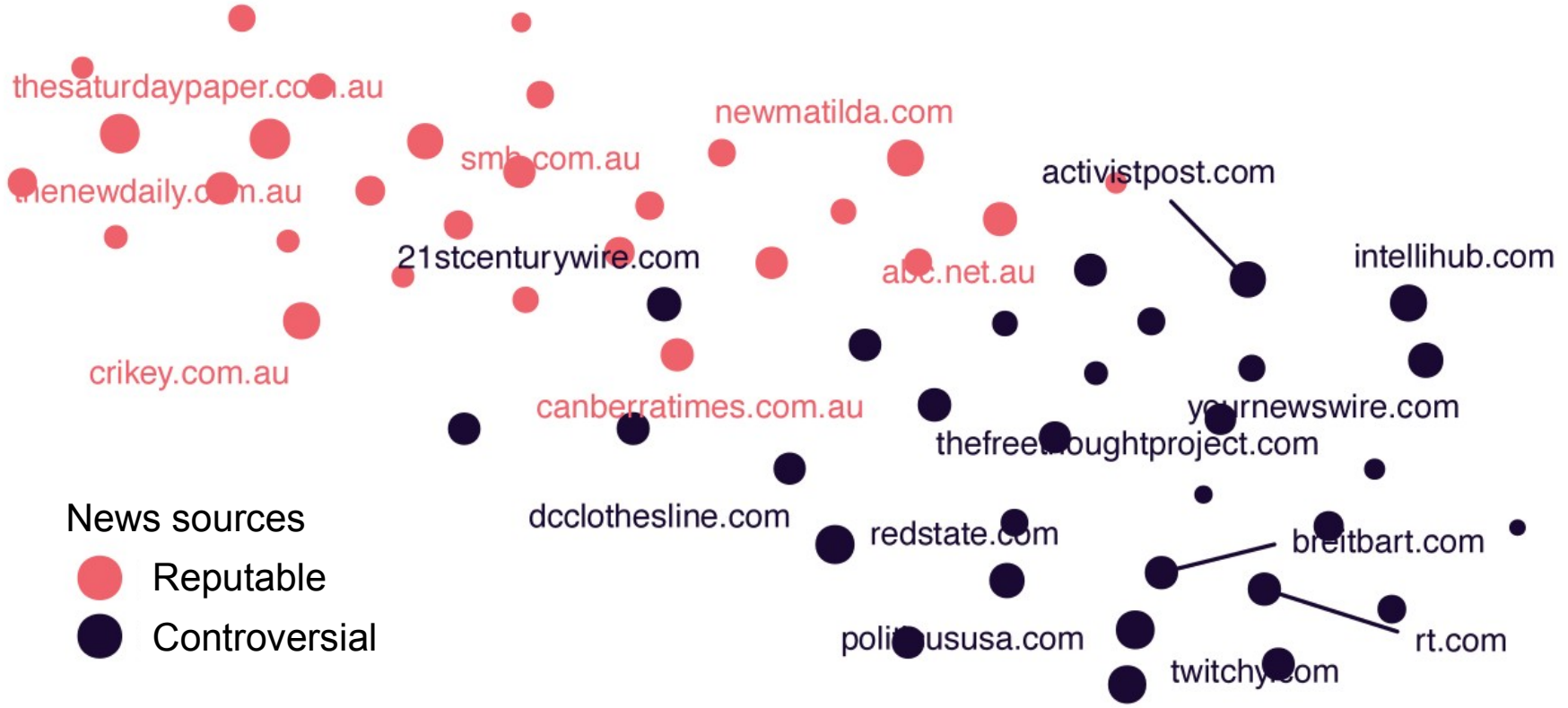
C – (news trolls) News about violence and civil unrest *#news*

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

The technical detail:

Semantic edit distance; dimensionality reduction; Twitter trolls

D. Expertise: Separating controversial from reputable



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

Detect controversial news without content analysis

The technical detail:

Modelling information cascades using mathematical generative models; Hawkes processes; joint modelling

Active UTS Research in the Disinformation Space

Highly collaborative cross-disciplinary research that connects cutting-edge data scientists with policy makers, social scientists, journalists, and defence organisations to deliver practical new ways of identifying, understanding and controlling problematic online content.

Objective

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

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

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?

Research

How do we detect disinformation?

What amplifies and widens the impact of disinformation?

How do we counter disinformation?

Approach

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

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

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

UTS Data Science Institute Capability

Team of 35 full-time data scientists and data engineers with a strong focus on delivering meaningful industry impact.

Long history of industry project delivery to diverse partners from Australian government, global water utilities, regulatory agencies, and energy, water, transport and education sectors

Deep expertise in cutting-edge social network message diffusion, virality and disinformation, ratified through high-profile publications

Data Science Institute members have won industry awards, the Eureka Data Science Prize, the CSIRO Collaboration medal for their work across applied data science initiatives

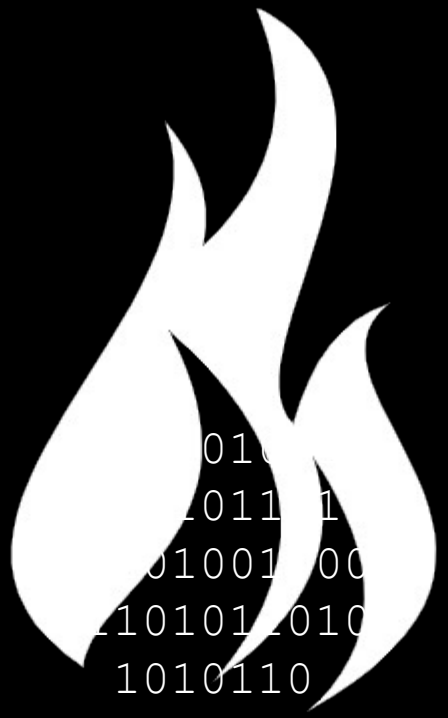
UTS is the top ranked university in computer science and engineering in Australia and top 15 in the world and is the top-ranked Australian university in scientific impact and collaboration

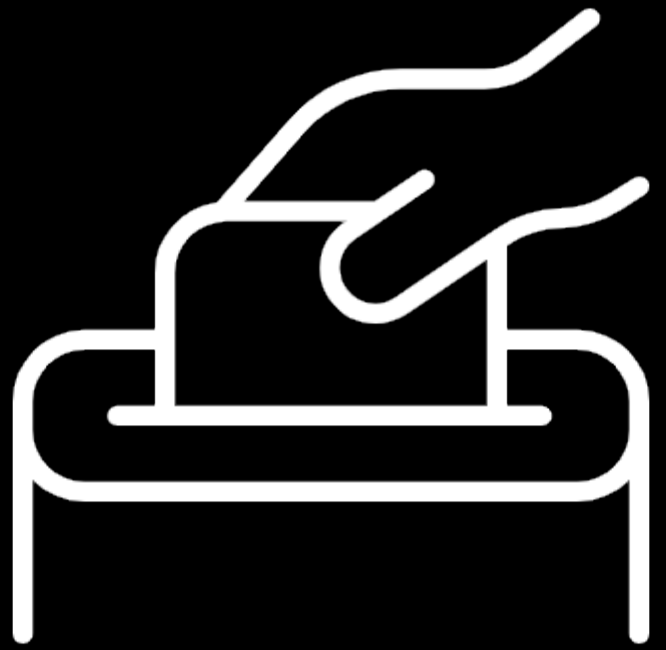
Experience in the management, leadership and delivery of large-scale collaborative research initiatives and long-term partnerships (including management of a $20m initiative with the federal government)

A collaborative network of researchers operating in the disinformation space and data science spaces, from PhD student through to senior researcher

A voice that stretches beyond academia, with meaningful media engagement record and experience across print, digital, television and radio platforms

Appendix:
context and technical details

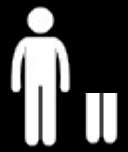




Disinformation wildfires at the state level

Disinformation in the United States

Internet Research Agency (IRA) disinformation attack on 2016 US presidential election



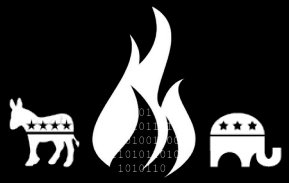
126 million Facebook accounts reached

187 million engagements on Instagram

288 million views on Twitter

For the 2020 US election, things are looking even worse

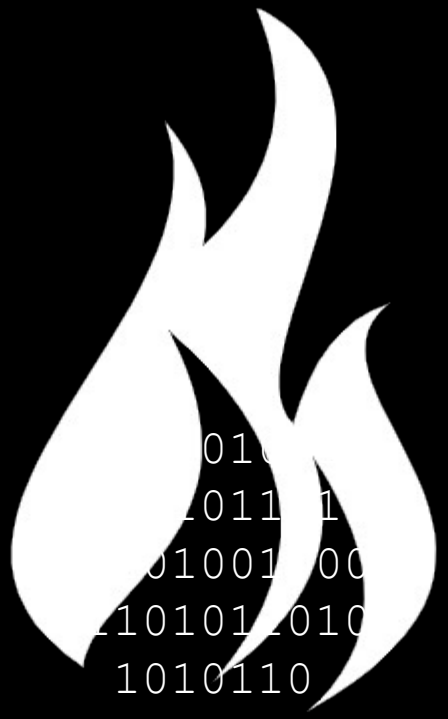
158.9 million Facebook accounts reached from just the top 100 politically-relevant fake news stories in 2019 alone



More than twice as many views of fake news stories on Facebook than for official Democratic and Republican party pages combined

2019 Worldwide Global Threat Assessment report from the US Director of National Intelligence

"For years, [our adversaries] have conducted cyber espionage to collect intelligence and targeted our critical infrastructure to hold it at risk. They are now becoming more adept at using social media to alter how we think, behave, and decide."



International governmental response

EU delivers the European Commission Action Plan against Disinformation, establishes the rapid alert system, and creates the strategic communication unit EUvsDisinfo

The National Defense Authorisation Act expands the Global Engagement Center's mission to include countering state-sponsored propaganda and disinformation, supported with an additional investment of $40m

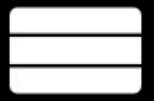
DARPA invests $68m to pursue the automated identification of manipulated videos and launches the SemaFor program on disinformation

France delivers the French Act against Informational Manipulation & iniates a co-regulatory engagement between government and facebook

Indonesia establishes the National Cyber and Encryption Agency with a primary focus on disinformation

Canada launches Critical Election Incident protocol and invests $7 million in dinsinformation awareness raising

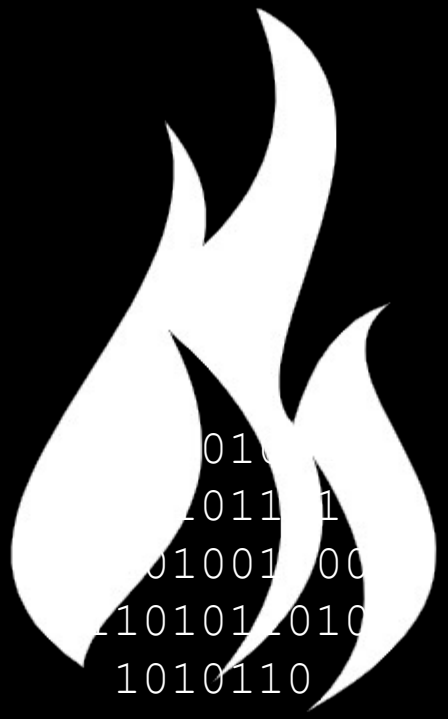
UK Government establishes and commits to ongoing funding for the (disinformation) Rapid Response Unit and invests £18 million on a 'fake news fund' for Eastern Europe

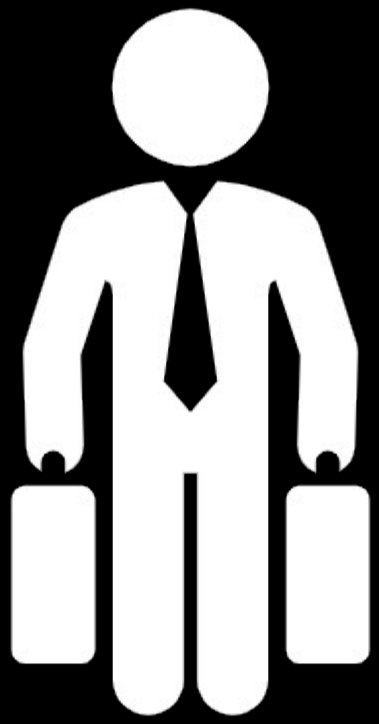


Germany launches the Network Enforcement Act and establishes Cyber and Information Domain Service to tackle cyber, foreign interference and disinformation threats

NATO's East Stratcom Taskforce provides monitoring of bot activity and disruption activities and produces the Disinformation Review

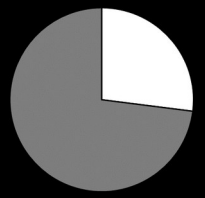
Google, Twitter, Facebook and Mozilla sign onto the EU Voluntary Code of Practice targeting the spread of disinformation



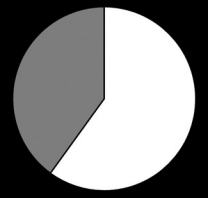


Disinformation wildfires at the market level

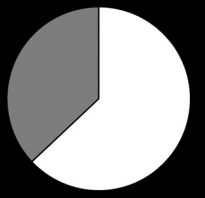
A survey of 588 large companies from across 13 countries underlines growing market concerns



27% report having already been significantly affected by adversarial social media



60% report concern about market manipulation



63% list adversarial social media activity as a priority risk area

The real impact of disinformation on the market is beginning to emerge

$4b

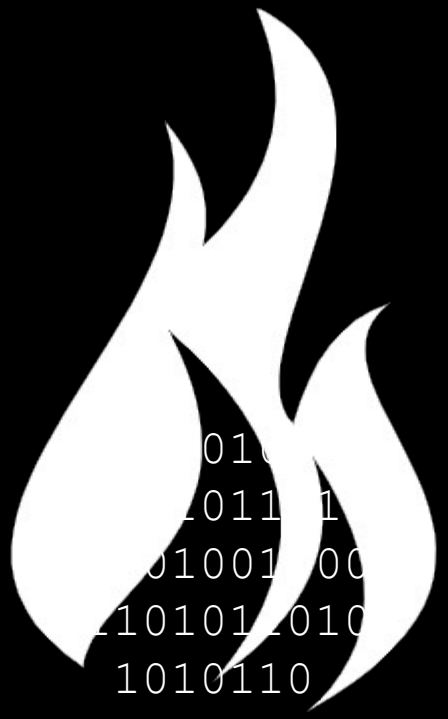
“Fake news stories in June 2017 reported Ethereum’s founder had died in a car crash and the company's market value dropped by $4 billion”

11%

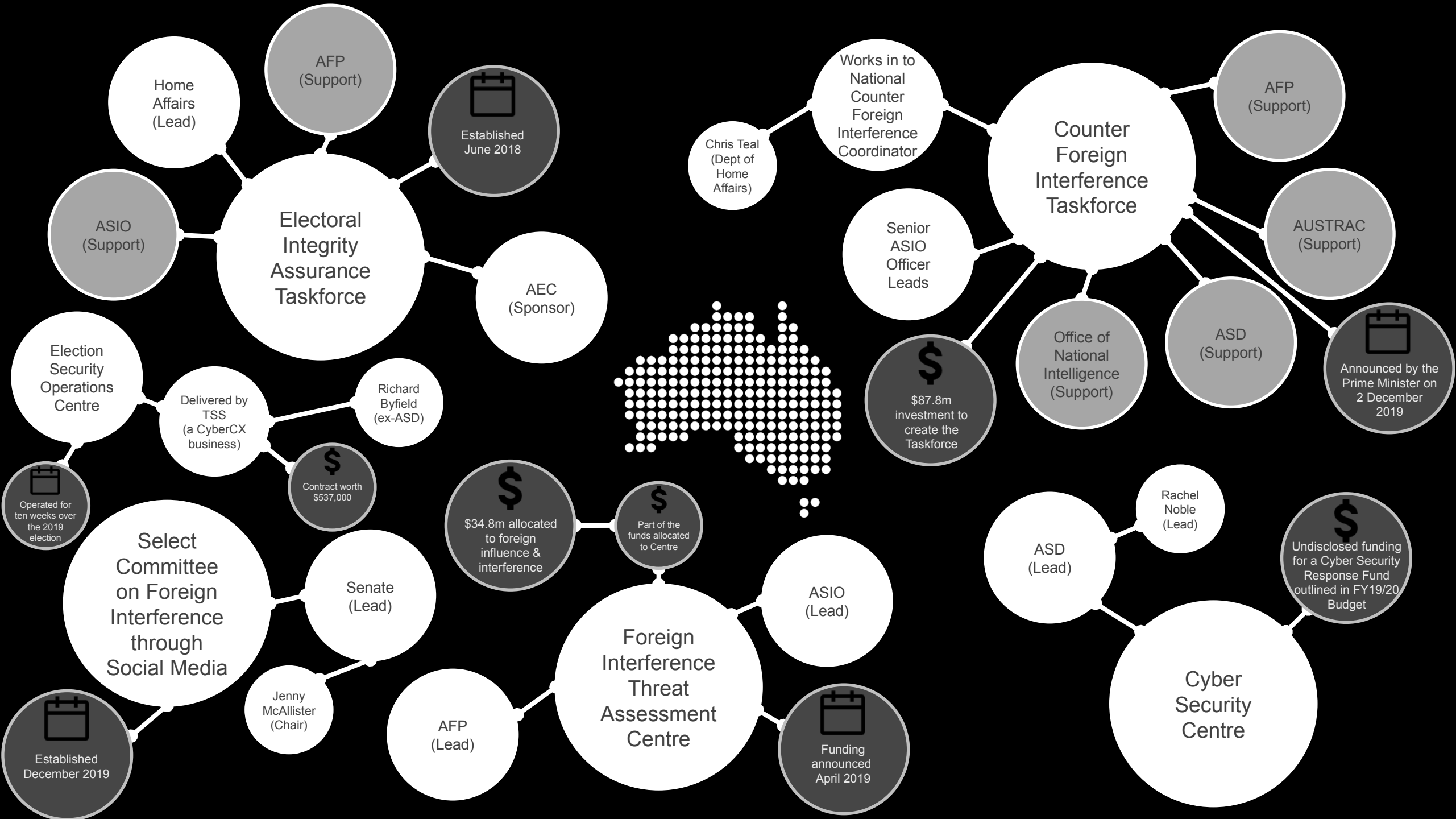
“Shares in the UK's Metro Bank plunged 11% before it could shake off inaccurate social media rumors that it was facing financial difficulties”

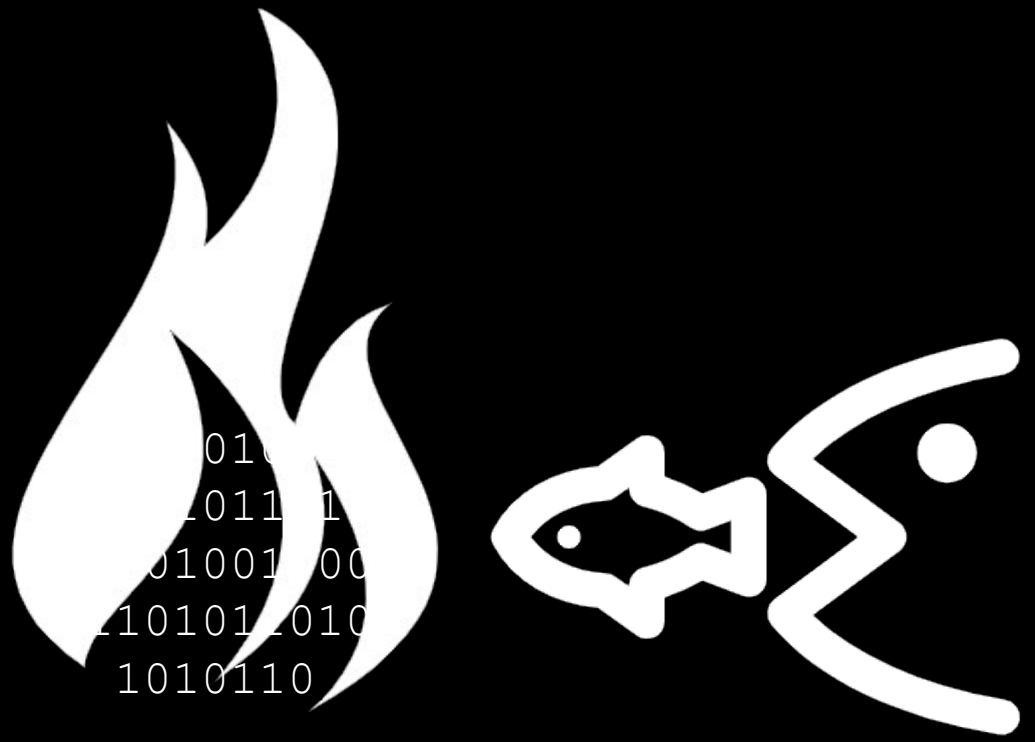
$39b

“…based on an analysis of past cases involving fake news inflicting damage on global stock markets, we find a potential [annual] loss of up to… $39 billion…”



Governmental response in Australia


Home Affairs (Lead)
AFP (Support)
Established June 2018
ASIO (Support)
Electoral Integrity Assurance Taskforce
AEC (Sponsor)
Election Security Operations Centre
Delivered by TSS (a CyberCX business)
Richard Byfield (ex-ASD)
Operated for ten weeks over the 2019 election
Contract worth $537,000
Chris Teal (Dept of Home Affairs)
Works in to National Counter Foreign Interference Coordinator
Counter Foreign Interference Taskforce
AFP (Support)
Senior ASIO Officer Leads
AUSTRAC (Support)
Office of National Intelligence (Support)
ASD (Support)
$87.8m investment to create the Taskforce
Announced by the Prime Minister on 2 December 2019
$34.8m allocated to foreign influence & interference
Part of the funds allocated to Centre
Select Committee on Foreign Interference through Social Media
Senate (Lead)
Jenny McAllister (Chair)
Established December 2019
AFP (Lead)
Foreign Interference Threat Assessment Centre
ASIO (Lead)
Funding announced April 2019
ASD (Lead)
Rachel Noble (Lead)
Undisclosed funding for a Cyber Security Response Fund outlined in FY19/20 Budget
Cyber Security Centre



A snapshot of the competitor landscape

Significant research programs in disinformation defence

- Misinformation Resilient Societies
  ~€4.11m from European Commission
- US govt provides $5m in total funding ($500k per project) via the Information Access Fund

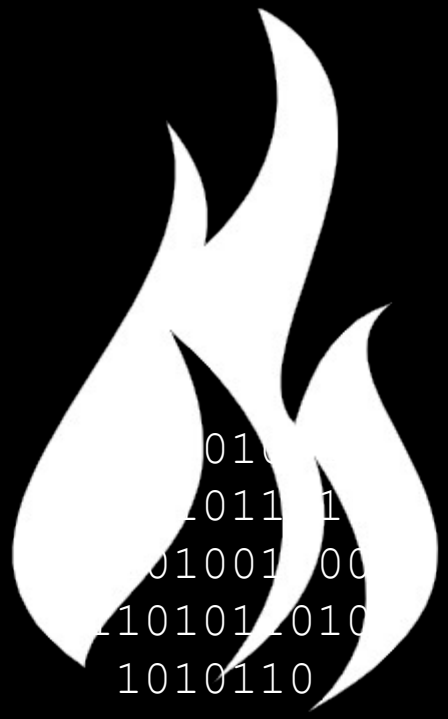
Start-ups

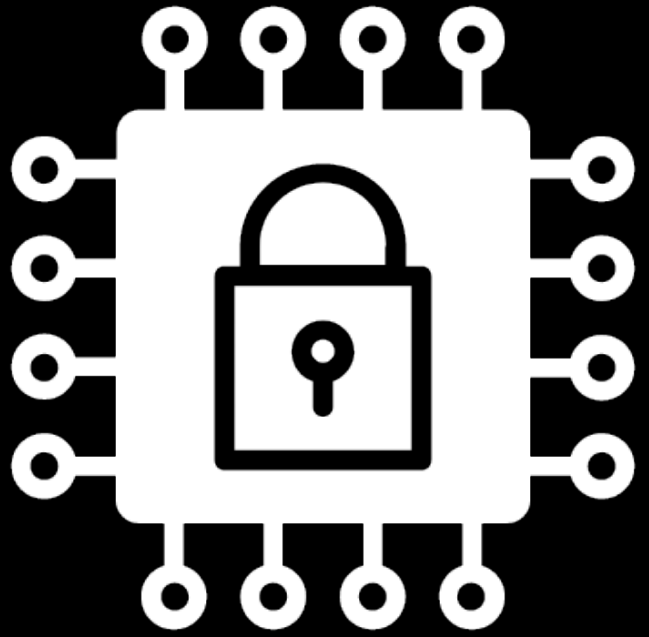
Predominantly fact-checking technologies

- NewsGuard
  ~$6m in VC funding
- CIVIL
  ~$5m in VC funding
- Factmata
  ~$1m in VC funding

Young companies providing disinformation defence services

- Digital Shadows
  ~$48m in VC funding
- Yonder
  ~$16m in VC funding



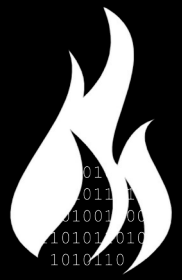


Is disinformation a cyber security problem?

Traditional cyber security failures are kindling to the fires of disinformation; a complete solution requires cyber and informational defences

Authenticity and reputation of message source

Disinformation delivery from hijacked accounts/ sites amplifies apparent authenticity and increases likely network attention



Potential impact and spread of disinformation

Disinformation delivery within compromised intranets/corporate communication platforms accelerates spread and increases likelihood of local impact

Network size, connectivity and communication viscosity

Message rhetoric, form and specificity

Stolen corporate/government data can be used for micro-targeting of disinformation campaigns and/or to underpin misleading narratives (where factual information is presented in incorrect contexts to maximise reputational damage)

"User access control, strong passwords, mandated multi-factor authentication, fraud detection, and identity theft prevention were just some of the **cybersecurity best practices that failed to combat Russian disinformation** just as much as fact-checking mechanisms or counter-narrative strategies."
RealClear Defense, 2018

Scott Jones, head of the Canadian Centre for Cyber Security: "**I never thought a cyber intelligence agency head would be talking about fake news, but it's something we have to tackle**."
Sydney Morning Herald, 2019

"…whereas [China and Russia] have recognized and developed mechanisms to **leverage the convergence of information warfare and cybersecurity**, America still tends to think of these activities as distinct and separate behaviors. We tend to think of our cyberdefenses as physical barricades, barring access from would-be perpetrators, and of information campaigns as retrograde and ineffective. In other words, we continue to focus on the walls of the castle, while our enemies are devising methods to poison the air." link
American Enterprise Institute, 2017

"On February 16, 2018, U.S. Department of Justice Special Counsel Robert Mueller indicted 13 Russians for interfering in the 2016 United States presidential election [1]... this indictment should make one thing clear: **information warfare is a cybersecurity issue**."
RealClear Defense, 2018

"Disinformation, by contrast, is a **new breed of a cybersecurity threat**… It's time to change how we think about propaganda, disinformation, and false information: it's not about fake news, **it's an adversarial attack in the information space**."
Yonder (disinformation solution provider), 2018

"The more legitimate an account appears, the more likely that the message will get amplified.
A compromised account…is the "perfect seed".
Decipher, 2019

Solutions

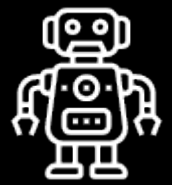
Contemporary solutions

Fact checking

- Manual fact checking — ⊗ Does not scale; significant human resourcing requirement; slow
- Automated fact checking — ⊗ Red Queen effect*
- Watermarking — ⊗ Requires widespread large-scale buy-in across content generators and platforms
- Journalism — ⊗ Does not scale; significant human resourcing requirement; vested interests; slow

Message source validation

- Block chain — ⊗ Computational complexity; requires widespread large-scale buy-in
- Journalism — ⊗ Does not scale; significant human resourcing requirement; vested interests; slow
- Watermarking — ⊗ Requires widespread large-scale buy-in; prone to falsification/Red Queen* effects
- Account authentication — ⊗ Prone to account hijacking



Bot identification

- Behavioural analysis — ⊗ Prone to account hijacking; Red Queen effect*; prone to hybrid human + bot solutions
- Content analysis — ⊗ Red Queen effect; unlikely to capture message amplification/recontextualisation

Platform control

- Content moderation — ⊗ Does not scale; significant human resourcing requirement; vested interests; slow
- Terms of use — ⊗ Vested interests; disinformation definitional difficulties

*The Red Queen effect refers to Alice in Wonderland, where the Red Queen says “Now, here, you see, it takes all the running you can do, to keep in the same place”. In contemporary artificial intelligence, it refers to the fact that, in an adversarial system, once a defensive mechanism is deployed, it can be used to train new AI solutions that are immune to that defence. This may lead to an endless unwinnable race for defenders.

Our solution

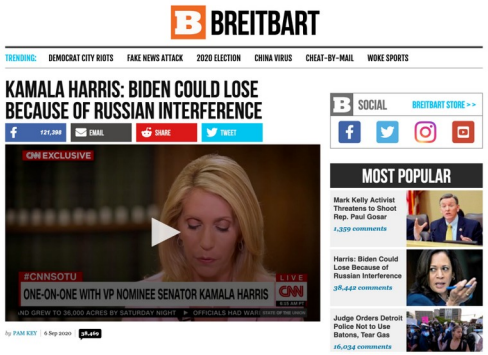
Don’t focus on the message; focus on the reaction

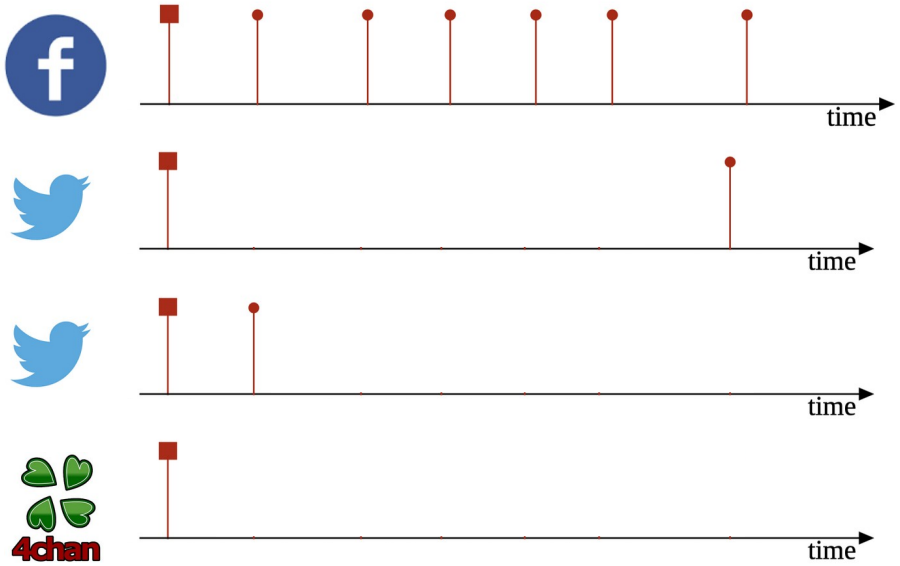
Quickly identify and prioritise high-risk content based on the way it is being consumed and communicated within and across communities. It is the *shape* of that communication that best tells us whether content will go viral, where it will spread next, whether we should be worried about its reach, and whether it displays the hallmarks of dangerous disinformation.

Capitalise on cutting-edge UTS Data Science Institute research into information diffusion, engagement and virality in social networks.

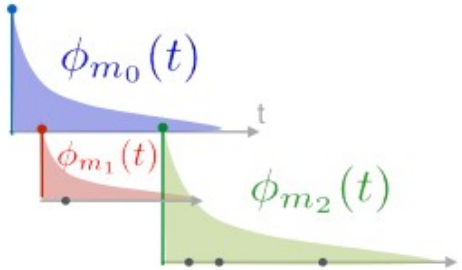
- Resilient to Red Queen effects, since risk assessment is based on population dynamics that are profoundly difficult for an attacker to control
- Not dependent on water-marking or source validation, removing need for content generator buy-in
- Computational approach is scalable, fast and does not carry significant human resourcing overhead

Expertise: Modeling diffusion cascades





Hawkes Process [Hawkes '71]



Our published work [Kong et al, CIKM'20]

Jointly model all cascades → **item-level diffusion modeling**;

Build **diffusion embeddings** usable with off-the-shelf classifiers

Kong, Quyu, Marian-Andrei Rizoiu, and Lexing Xie. 2020. "Describing and Predicting Online Items with Resharing Cascades via Dual Mixture Self-Exciting Processes." in ACM International Conference on Information and Knowledge Management (CIKM'20).